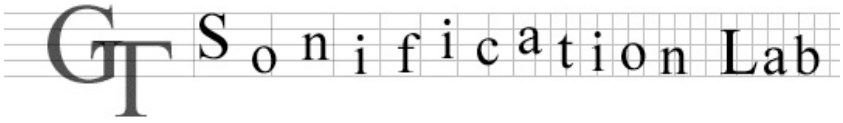
ICAD 2012 | The 18th Annual International Conference on Auditory Display
Atlanta, GA 18th ~ 21st June 2012

PROCEEDINGS OF THE 18TH INTERNATIONAL CONFERENCE ON AUDITORY DISPLAY

June 18-21, 2012, Atlanta, GA, USA

Edited by
Michael A. Nees
Bruce N. Walker
Jason Freeman



PUBLISHED BY:
The International Community for Auditory Display

CREDITS:
Logo design: Rick Swette

ICAD 2012 homepage: www.icad2012.icad.org

ISSN: Forthcoming

TABLE OF CONTENTS

Preface

Papers

Posters

Extended Abstracts

Sonification Contest Finalists

Organizing Committee

General Chair
Bruce N. Walker, Georgia Institute of Technology

Program Chair
Michael A. Nees, Lafayette College

Music Chair
Jason Freeman, Georgia Institute of Technology

Doctoral Consortium and Student ThinkTank Chair
Terri L. Bonebright, DePauw University

Workshops Chair
Tae Hong Park, Georgia State University

Accessibility Chair
Carrie Bruce, Georgia Institute of Technology

Webmaster
Rick Swette, Georgia Institute of Technology

Proceedings Peer Reviewers

Kovacs Balazs
Michael Banf
Stephen Barrass
Jared Batterman
Terri L. Bonebright
Jeffrey Boyd
William Brent
Derek Brock
Carrie Bruce
Hans-Peter Brueckner
Marcelo Caetano
Hongchan Choi
Perry R. Cook
Benjamin Davison
Liz Diaz
Milena Droumeva
Marc Anders Ericson
Samuel Ferguson
John Flowers
Christopher Frauenberger
Jessica Gillard
Christian Gonzalez
Joachim Gossman
Matti Grohn
Florian Grond
Paula Henry
Norbert Herber
Thomas Hermann
Andy Hunt
Myounghoon Jeon
Savvas Eddy Kazazis
Permagnus Lindborg
Jeffrey Lindsay
Vincent Martin
David Mauro
Ryan McGee
Julia D. Olsheski
Neel S. Patel
S. Camille Peres
Rhys Perkins
David Pirro
Agnieszka Roginska
Nina Schaffert
Angelique Scharine
Margaret Schedel
Gerd Schmitz
Timothy Schmele
Jonathan Schuett
Raymond M. Stanley
Tony Stockman
Alexandra Supper
Paul Vickers
Katharina Vogt
Marcus Watson
Gyorgyi Wersenyi
David Worrall

Sonification Contest Jury

Alberto de Campo
R. Luke Dubois
Adam Lindsay
Brian Whitman

Keynote Addresses

REFLECTIONS OF A GIRL AUDIO GEEK
June 18, 2012
Elizabeth Mynatt
Professor of Interactive Computing, Executive Director of Georgia Tech's Institute for People and Technology.

In 1992, I had the tremendous opportunity to participate in the first gathering of the International Community for Auditory Display. Thirty-five researchers engaged in focused discussions and communal laughter as we sought to build a community dedicated to understanding the experience of sound. The community was a good fit for me as the experience of sound had been a great teacher, informing my research and my life as a researcher. By working with sound, and by working with people who's dominant experience of the world was sound, I gained an early appreciation for sound as the core of experience design. Through sound we tell stories, we convey emotion and we convey the fundamental qualities of places. These lessons have journeyed with me over the past 20 years and have informed much more than the design of auditory displays. Sound is a great teacher and I have been privileged to be its student.

SURPRISE IN SONIFICATION AND SONIFYING SURPRISE
June 19, 2012
Jonathan Berger
Denning Provostial Professor in Music, The Center for Computer Research in Music and Acoustics (CCRMA), Stanford University

Along with timbre, perhaps the most critical components of effective auditory display is the use (or abuse) of expectation realization and violation. Yet, remarkably, (again, along with timbre), to date there exists neither quantitative metric nor qualitative descriptors for expectation. In this talk I will present the formulation and control of expectation as a principle attribute in the creation and experience of both music and auditory display.

Acknowledgments

The organizers of ICAD 2012 gratefully acknowledge the contributions of the following people to the success of the conference:

Derek Brock served as the conflict-of-interest editor during the peer-review of submissions from the program committee.

Jonathan Schuett, Thom Gable, and Ben Davison coordinated logistics for the conference.

Carrie Bruce arranged and coordinated the conference banquet.

Julia Olsheski arranged the welcome reception.

Katie Gentilello and Julie Speer provided valuable support and assistance with the conference website and the electronic archiving of the proceedings

The ICAD Board generously provided a loan for initial conference expenses and funded the monetary awards for best submissions.

Derek Brock, Carrie Bruce, Milena Droumeva, Matti Grohn, Thomas Hermann, Myounghoon Jeon, Camille Peres, and Nina Schaffert chaired sessions of presentations during the conference.

Jared Batterman coordinated food and refreshments during the conference.

Derek Brock, Terri Bonebright, Perry Cook, Myounghoon Jeon, Camille Peres, and Tae Hong Park conducted conference workshops.

Sereatha Hopkins and Kristin Pealer provided support with processing payments and budgeting.

Shawn Stinson was the event coordinator at the Academy of Medicine.

Adam Lindsay and Social Genius provided the Twitter Music Trends API used in the sonification competition.

Frank Clark and the Georgia Tech School of Music provided audio equipment for paper presentations, sonification contest finalists, and the concert.

Tom Sherwood and Jessica Peek Sherwood from Sonic Generator helped to organize the concert.

AN ALTERNATIVE IMPLEMENTATION OF VBAP WITH GRAPHICAL INTERFACE FOR SOUND MOTION DESIGN

Hongchan Choi

Stanford University
Center for Computer Research in Music and Acoustics (CCRMA)
660 Lomita Dr, Stanford, California, USA
hongchan@ccrma.stanford.edu

ABSTRACT

An implementation of vector-based amplitude panning (VBAP) for spatial display of sonified data is presented. The proposed method offers an implicit conversion from Spherical to Cartesian coordinates thus being particularly well suited for auditory display. Two techniques from computer graphics are adapted in order to predefine an optimum set of speaker triplets and perform the amplitude panning in real-time. Furthermore, the consideration of time delay from a virtual sound source to actual speakers is incorporated. Due to the geometrical nature of this procedure, the resulting system can be easily visualized by the graphic library OpenGL. Using this library I provide users with an intuitive control interface. A prototype is demonstrated that enables a user to compose a trajectory of sound in three dimensional space.

1. MOTIVATION

1.1. VBAP: Vector-Based Amplitude Panning

Vector-based amplitude panning (VBAP) [1] is one of several non-standard methods used to render virtual sound sources in 3D sound field using multiple speakers. VBAP is distinct in its clustering of adjacent speakers into triplets in which individual gain factors are calculated for each speaker in order to translate the sound into perceptually compelling spatial auditory cues.

The conventional VBAP method includes the following steps:

a) Define speaker triplets.

b) Position a virtual sound source (a new vector P) in the space.

c) Select a triangle (a speaker triplet) intersected by a vector between a sound source (P) and the position of listener (L).

d) Calculate 3 gain factors from each speaker on the triplet.

e) Interpolate gain factors from previous ones to new ones.

f) Iterate through steps b - e as needed.

Although a few implementations of VBAP have been adapted since the method's introduction in 1997 [2] [3], the procedure described here substitutes steps a), c) and d) with techniques from computer graphics. A novel approach emerges with a more intuitive visual interface and enhanced computational efficiency. The prototype transforms a spherical system (ambisonics and VBAP) into the Cartesian coordinate system, the one used by standard graphic libraries. This transformation bears a number of significant advantages including:

- integration with conventional graphic libraries, such as openGL and 3D vector calculation,
- a mapping paradigm that integrates well with data visualizations,
- an intuitive means of positioning and moving sound in virtual 3D space.

We describe Implementation of two algorithms adapted from 3D graphics, Quickhull [4] and Ray-Triangle intersection [5] following a description of *Field 8* a multi-touch interface for 2D panning.

1.2. *Field 8*, a multi-touch interface for 2D panning based on DBAP

Field 8 is a control user interface designed for distance based amplitude panning (DBAP) [6]. It provides an intuitive control interface on a multi-touch screen implemented on the iPad . Unlike other user interfaces of sound spatialization for VBAP or Ambisonics [7], the real-time user interaction and the rich visual feedback are focal points of the interface that allows users to draw multiple paths of sound motion with up to eight fingers. The site-specific prototype design for the CCRMA listening room [8], encompassing the user interface on iPad and a spatialization server built with ChucK audio programming language [9],provided a suitable environment to explore the sonic space in a 2D plane created by 8 speakers at ear level. The prototype has been successfully used in various performance contexts and compositions. *Field 8* is also useful for exploration and rapid design of appropriate scaling and mapping methods for auditory display. Expanding this potential to 3D auditory display using more than 8 speakers is clearly the next step.

1.3. Considering Efficiency and Usability

Adapting *Field 8* to 3D space using more than 8 speakers presents a number of logistical problems. The algorithm for DBAP should be redesigned to update an arbitrary number of gain controllers (in the site specific case cited here, 22 gain controllers) every few milliseconds. Deploying multiple sound sources will multiply the number of gain controllers. For example, a DBAP system requires to update 176 gain factors for every sample when there are 8 sound sources in motion. Furthermore, the system must calculate 176 distances between 3D points to get each gain factor meaning that the process involves 176 square-root operations in every iteration. A more efficient approach was needed to build a real-time spatialization system capable of handling multiple user interaction. The

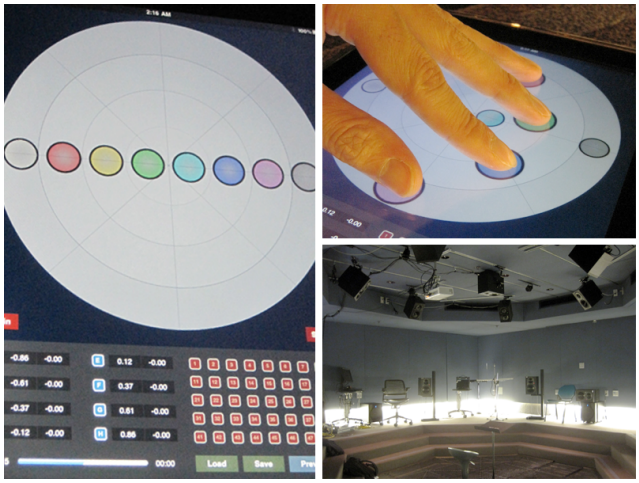


Figure 1: User interface of *Field 8* and CCRMA listening room

VBAP concept of grouping three proximate speakers proved a viable option which achieved a decent level of interactivity.

To summarize, the motivation for this work is twofold: developing a new panning system that

- can move multiple sound sources in a highly efficient fashion, and
- enables the user to design sound motions in an intuitive and expressive way.

2. SYSTEM DESIGN

2.1. Phase 1: Triangulation of Speakers

The operation of VBAP includes two separate procedures; the first phase is organizing the physical position of speakers in the space [2]. This procedure obtains an optimum set of non-overlapping speaker triplets(triangles), thus reducing the computational load by dealing with only 3 speakers at any given moment to calculate the panning. Unless the physical configuration of speakers is changed, the first phase needs to be updated only once.

Triangulating contiguous speakers in 3D space can be done in several different ways. Although the original source code uses triangulation [10] there was no specification in the original VBAP algorithm and subsequent papers of how it was implemented. My approach on this grouping task is to use a convex hull algorithm called Quickhull3D [4]. The convex hull of a set of points is the smallest convex set that contains the points. The algorithm originated from the field of computer graphics and is widely used to build a 3D mesh with a minimum set of triangles from an arbitrary number of vertices. (See figure 2.)

The Quickhull algorithm is highly optimized, so inserting a new vertex into the existing 3D mesh to build a new set of triangles on the fly is possible. Considering the largest speaker system in the world is using less than 200 speakers and the Quickhull algorithm can handle more than 200 vertices in real-time fashion, this algorithm is a viable way to triangulate speakers.

2.2. Phase 2: Ray-Triangle Intersection

The original VBAP algorithm calculates gain factors from a selected triplet by matrix operation. However, another method from computer graphics can be deployed to get gain factors. The Ray-Triangle intersection algorithm [5] is a 3D vector operation to calculate not only intersection of a ray vector and a triangle, but also the point of intersection. (See figure 3.)

If the position of a virtual sound source exists as a 3D point in the space, than we can assume a vector from the point of origin to the point of the sound source. If the physical configuration of speakers is a spherical mesh of triangulated speakers, the infinite extension of this vector intersects only one triangle (speaker group). This is particularly useful for VBAP operation because the Ray-Triangle intersection algorithm can infer an intersection point on a triangle, and the distances between 3 speakers of the triangle and this point can be calculated.

2.3. Relative Loudness and Time Delay

The Ray-Triangle intersection algorithm yields useful parameters. Rendering a realistic 3D auditory display is possible by using the loudness ratio between 3 speakers as well as the time delay estimated from the distance between a sound source and the speakers. Parameters from the algorithm are listed here (see also Figure 4).

a) A gain factor estimated from the distance between a sound source and the listener: P-L (in figure 4-(a))

b) 3 loudness ratios from distances between 3 speakers and a intersection point: S0-I, S1-I, S2-I (in figure 4-(b))

c) 3 delay times estimated from distances between 3 speakers and a sound source: S0-P, S1-P, S2-P (in figure 4-(c))

The system yields gain factors for 3 speakers by summing all 3 distances and dividing each distance by the sum as described in b). For example, when the intersection point moves to the exact same

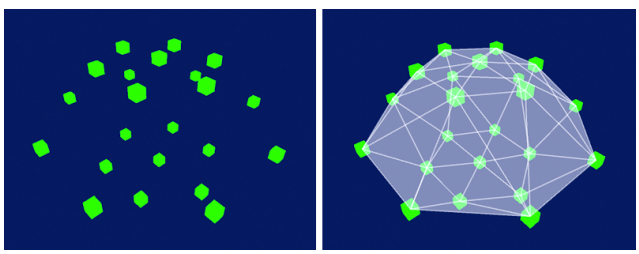


Figure 2: Using Quickhull algorithm to triangulate speakers

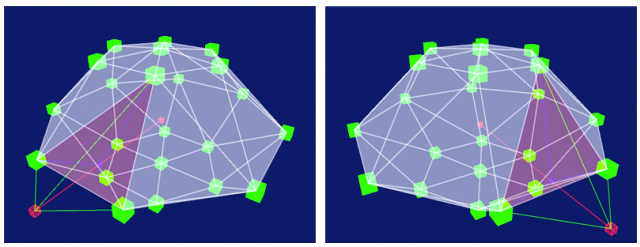
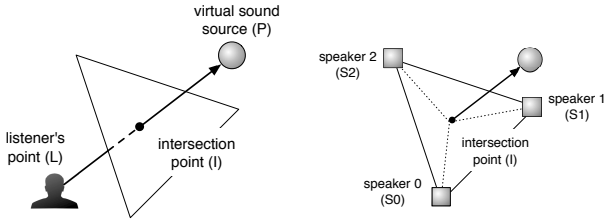
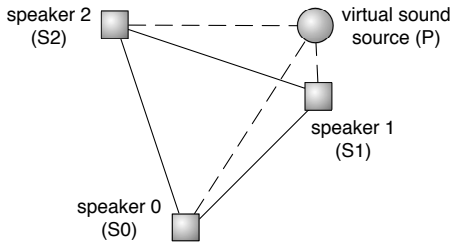


Figure 3: Ray-Triangle Intersection algorithm to select a triplet



(a) intersection point

(b) distance from 3 speakers



(c) distance from source

Figure 4: various relationships from intersecting algorithm

location as one of the 3 speakers, the gain factor for the speaker becomes 1.0 reducing the gain factors of the other two speakers to zero. In most cases, the intersection point moves from one triangle to another by passing through their shared edge. When the intersection point is located precisely on the edge, the sum of relative loudness of the two speakers on that edge will be 1.0. This will ensure the seamless transition when the intersection point moves across two triangles. This is partly similar with the DBAP method; however, it differs in its use of 3 speakers at any given moment. Also the distance between the listener and the sound source affects the overall loudness of the sound.

The original implementation of VBAP lacks the notion of time delay between a virtual sound source to selected speakers. Simply by calculating distances between a sound source from speakers in a selected group, the system can simulate time delay introducing the subtle change of timbre that arises from phase differences. Such concepts are integral to Wave field Synthesis.

Here we encounter a common obstacle in artificial spatialization methods: When the position of a virtual sound source is inside of a sound field the time delay of speakers will be a negative value, which is impossible in the real world, causing ambiguity in localization of the sound. Thus, this problem still remains in the system.

3. IMPLEMENTATION OF PROTOTYPE

A prototype built with two programming languages, Processing and ChucK, is demonstrated to test the feasibility and possible enhancements. Processing [11] functions as a core system that calculates the entire panning process and sends the result to ChucK [12] via an OSC(OpenSound Control) [13] connection. The VBAP object in ChucK renders a sound source in the space according to the data from a speaker triplet delivered from Processing. In this section, I describe design choices and details on implementation.

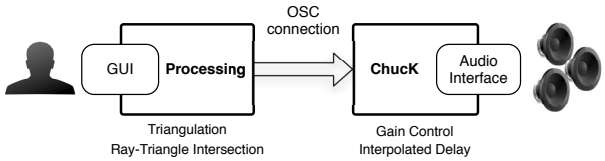


Figure 5: 2-Tier structure: Processing and ChucK via OSC

3.1. Processing: Spatialization Engine and User Interface

Processing is a visual programming language built for media arts and the classroom setting. It is widely used for designing prototypes or creating visual arts. One of its benefits is a large set of libraries that can be deployed with minimum effort. It is especially helpful to visualize data structures for better understanding.

The prototype implements Processing mainly because it has a nice library of vector calculation and a built-in OpenGL support. It significantly cuts the development time. A visualization is inherently correlated with a graphical user interface; thus having a compelling visualization is a clear advantage in terms of user control.

Since actual positions of speakers were initially in a spherical coordinate system, typical of a spatial audio setting, converting them into Cartesian coordinate system was required. This can be achieved when we understand that Zenith in a normal spherical system and Elevation in spatial audio are two orthogonal descriptions a vertical angle. This conversion into a Cartesian system enables us to perform a vector operation, a great advantage in terms of not only visualizing or animating what is happening, but also calculating required values from geometry algorithms since standard graphic libraries are based on the Cartesian system.

The system reads the speaker position data, converts them into Cartesian values, and then performs the quickHull3D algorithm to get an optimum set of speaker triplets. Convex hulling is a type of triangulation algorithm, that differs from the original triangulation algorithm in VBAP. In my implementation, this step is done by a library called newHull, a ported library from the original Java

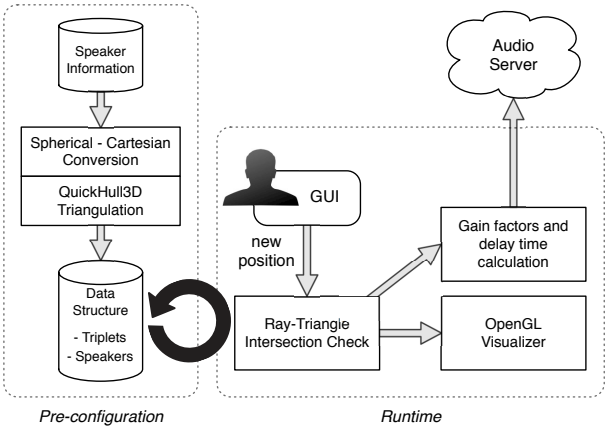


Figure 6: Sptialization Engine in Processing

implementation of quickHull3D.

The quickHull object in Processing is designed to produce a set of vertices(speakers), face indices (triangles, speaker triplets).

As a next step, the Ray-Triangle intersection algorithm checks all the triplets with a vector between the origin(listener) and a sound source. (See figure 4 (a).) The function that contains The Ray-Triangle intersection algorithm is the core of the whole system. It performs not only the essential visualization (lines between selected speakers and a sound source) but also calculates all the gain factors with delay times and sends OSC messages to the audio server. The OSC message consists of 3 parameters: an ID of a speaker, a gain factor(zero to one), and a delay time in milliseconds.

Moving a mouse can control the position of sound source. The camera will gradually follow the position of the sound source as it moves. Unlike other conversions made in the system, the movement of the mouse in a 2D plane can be converted into a 2 angle (Zenith, Azimuth) spherical coordinate system.

3.2. ChucK: Multi-Channel Audio Server

ChucK is a general-purpose programming language tailored for computer music. [12] miniAudicle, the front-end of the ChucK virtual machine, accelerated the prototype design process supporting concise programming and rapid experimentations. [14] A multichannel audio server is implemented in the ChucK language with miniAudicle. This server features 22 channels of audio to represent one or multiple sound sources in this iteration of the prototype. As mentioned, this prototype was designed for the CCRMA listening room. The site-specific details of this implementation are described in the next section.

The OSC data stream is dispatched to a respective speaker by the ID field. Note that the number of OSC packets is constantly 3 per speaker triplet and the Processing OSC sender will send these 9 numbers at every frame (about 16.7 milliseconds), 540 numbers per second. This is 7.3 times better than sending OSC packets for the entire set of 22 speakers with 3960 numbers for every second. For example, Field 8 was designed to control 8 speakers with 1440 OSC packets per second and occasionally encountered a bandwidth problem.

This is a significant advantage when detaching the control interface into a wireless device such as an iPad allowing users to control positions of sound sources without sitting in front of a workstation. The OSC packet size is consistently 9 numbers regardless of the number of speakers to be controlled and this consistency is possible thanks to the pre-configuration process of VBAP.

The interpolation between successive gain factors is performed by the built-in features of the Envelope objects in ChucK. The duration of interpolation is 16.7ms corresponding to the data speed from Processing rendering a seamless transition from previous gain factors to next ones. Unit generators for delay(DelayA) are interpolating delay times by default, so changing delay time does not introduce discontinuity in samples.

4. THE CCRMA LISTENING ROOM: SITE SPECIFIC SETUP

The CCRMA listening room is an experimental 3D space with 22 speakers and near-anechoic acoustics. The default 3D panning scheme is 3rd order Ambisonics (3v3h) that utilizes 16 channels of encoded audio streams. The OpenMixer [15] is a highly flexible software mixer running on the workstation. It transforms these encoded 16 streams into 22 audio channels routing the 22 speakers distributed in a sphere around the listener. The panning operation is accessible through a few experimental panners in Ardour [16], PureData [17] and SuperCollider [18].

As previously discussed, the prototype is tuned for the setting of the CCRMA listening room. However, it does not mean that the speaker position data is hard-coded in the software. Unlike the Ambisonics implementation, the new VBAP implementation performs triangulation on the fly from a text file with the positional information (either Spherical or Cartesian format). Therefore, it can be easily adapted to other venues without redesigning an encoder or a decoder.

The OpenMixer provides highly flexible audio input arrays including netJack [19] or JackTrip [20]. The ChucK audio server running on the laptop (MacBook Pro with a dual core CPU at

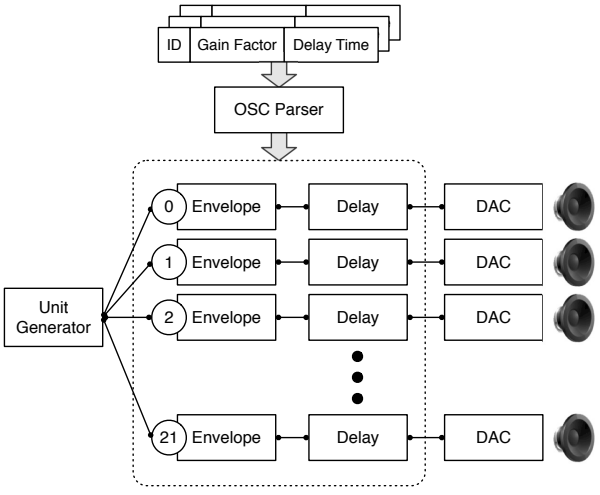


Figure 7: Multi-Channel Audio Server in ChucK

Figure 8: The CCRMA Listening Room

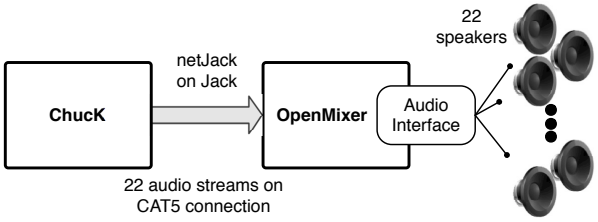


Figure 9: Connection between Audio Server and OpenMixer

2.2Ghz) could transfer 22 audio streams through the netJack driver without any problem.

5. CONCLUSION AND FUTURE WORK

In this study, I investigated a new VBAP implementation adopting two techniques from computer graphics to improve several aspects. The prototype presented the more intuitive and expressive graphical control interface. The embedded transformation of coordinate system creates synergies by tapping computer graphics libraries resulting in a highly responsive control interface.

However, the early implementation of the application has limitations. It does not have a sequencing feature yet, so it is not possible to record the trajectory of the sound. The system handles the sound material as a sound entity, rather than under the framework of "audio track" like typical sequencers or digital audio workstations. As of now, this prototype features only real-time interaction.

Future goals include a more sophisticated graphical user interface for 3D panning. This interface will include wireless and touchscreen devices for portability. To facilitate the calibration of varying speaker setups, I foresee using computer vision or 3d camera technology to quickly convey measurements to the system. Greater efficiency might be achieved if we integrate the panning operation into Chuck as a built-in unit generator. To mitigate the challenges of diversified setups, we could have a standardized method for notating speaker configurations. For example, the system could be calibrated for a given space by downloading an XML file from the website.

6. ACKNOWLEDGMENT

I would like to thank John Granzow and Jonathan Berger who provided invaluable suggestions and advice, to Fernando Lopez-Lezcano, for sharing his expertise in spatial audio and the CCRMA listening room. I am also grateful to Jarosalw Kapuscinski who provided with artistic and creative directions for the *Field 8* application and to Chris Chafe and Ge Wang whose related work inspired this research.

7. REFERENCES

[1] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *Journal of the Audio Engineering Society*, vol. 45, no. 6, pp. 456–466, 1997.

[2] V. Pulkki and T. Lokki, "Creating auditory displays to multiple loudspeakers using vbap: A case study with diva project," in *International Conference on Auditory Display*, 1998.

[3] V. Pulkki, "Generic panning tools for max/msp," in *Proceedings of International Computer Music Conference*, 2000, pp. 304–307.

[4] C. Barber, D. Dobkin, and H. Huhdanpaa, "The quickhull algorithm for convex hulls," *ACM Transactions on Mathematical Software (TOMS)*, vol. 22, no. 4, pp. 469–483, 1996.

[5] T. Möller and B. Trumbore, "Fast, minimum storage ray-triangle intersection," *Journal of graphics tools*, vol. 2, no. 1, pp. 21–28, 1997.

[6] T. Lossius, P. Baltazar, and T. de la Hogue, "Dbap-distance-based amplitude panning," in *Proceedings of 2009 International Computer Music Conference, Montreal, Canada*, no. 1, 2009.

[7] M. Gerzon, "Ambisonics in multichannel broadcasting and video," *J. Audio Eng. Soc*, vol. 33, no. 11, pp. 859–871, 1985.

[8] F. Lopez-Lezcano and C. Wilkerson, "Ccrma studio report," in *Proceedings of the International Computer Music Conference*, 2007.

[9] G. Wang, P. Cook *et al.*, "Chuck: A concurrent, on-the-fly audio programming language," in *Proceedings of International Computer Music Conference*, 2003, pp. 219–226.

[10] V. Pulkki, "Tkk akustiikka/tkk acoustics laboratory/software," http://www.acoustics.hut.fi/software/, retrieved February 2012.

[11] C. Reas and B. Fry, "Processing: programming for the media arts," *AI & Society*, vol. 20, no. 4, pp. 526–538, 2006.

[12] G. Wang, "The chuck audio programming language." a strongly-timed and on-the-fly environ/mentality"," Ph.D. dissertation, Princeton University, 2009.

[13] M. Wright, A. Freed, and A. Momeni, "Opensound control: State of the art 2003," in *Proceedings of the 2003 conference on New interfaces for musical expression*. National University of Singapore, 2003, pp. 153–160.

[14] S. Salazar, G. Wang, and P. Cook, "miniaudicle and chuck shell: New interfaces for chuck development and performance," in *Proceedings of the 2006 International Computer Music Conference*, 2006, pp. 63–66.

[15] F. Lopez-Lezcano and J. Sadural, "Openmixer: a routing mixer for multichannel studios," in *Linux Audio Conference 2010*, UtrechT, The Netherlands, 5/2010 2010.

[16] P. Davis *et al.*, "Ardour: digital audio workstation," http://ardour.org/, retrieved February 2012.

[17] T. Musil, M. Noisternig, and R. Höldrich, "A library for realtime 3d binaural sound reproduction in pure data (pd)," in *Proc. 8th Int. Conference on Digital Audio Effects*, 2005.

[18] J. McCartney, "Rethinking the computer music language: Supercollider," *Computer Music Journal*, vol. 26, no. 4, pp. 61–68, 2002.

[19] A. Carôt, T. Hohn, and C. Werner, "Netjackremote music collaboration with electronic sequencers on the internet," in *Proceedings of the Linux Audio Conference*, 2009.

[20] J. Cáceres and C. Chafe, "Jacktrip: Under the hood of an engine for network audio," *Journal of New Music Research*, vol. 39, no. 3, pp. 183–187, 2010.

SOUND EFFECT METAPHORS FOR NEAR FIELD DISTANCE SONIFICATION

Gaëtan Parseihian & Brian FG Katz

LIMSI-CNRS
BP133, 91403, Orsay, France
Gaetan.Parseihian@limsi.fr
Brian.Katz@limsi.fr

Simon Conan

LMA-CNRS
31 chemin Joseph-Aiguier
13402 Marseille, France
conan@lma.cnrs-mrs.fr

ABSTRACT

This article presents a concept of distance sound source sonification for virtual auditory displays in the context of the creation of an assistive device for the visually impaired. In order to respond to user needs, three sonification metaphors of distance based on sound effects were designed. These metaphors can be applied to any type of sound and thereby satisfy all aesthetic desires of users. The paper describes the motivation to use this new type of sonification based on sound effects, and proposes guidelines for the creation of these three metaphors. It then presents a user evaluation of these metaphors by 16 subjects through a near field sound localization experiment. The experiment included a simple binaural rendering condition in order to compare and quantify the contribution of each metaphor on the distance perception.

1. INTRODUCTION

Thanks to the development of research in auditory display, the use of sound as a means to convey information has considerably grown over the past few decades. One of the most obvious applications is the sensory substitution of visual information when it is not available. Visually impaired people have a variety of needs for non-visual information. Accessing computer information, avoiding obstacles, finding a route or a desired inanimate object are examples of tasks that can be challenging for them. Some of these problems could be resolved by the use of auditory displays.

This study takes place within the context of the development of an electronic device based on rapid object localization and auditory augmented reality for helping people with visual impairments in near field guidance (hand reaching movement for grasping objects) [1]. This device combines a bio-inspired vision system able to quickly recognize and locate objects [2] and a 3D sound rendering system [3] which will map a spatialized sound to the location of the targeted object. Sound guidance will be provided through binaural rendering, allowing a full exploitation of the human perceptual and cognitive capacity for spatial hearing.

Even though the basic mechanisms of directional sound localization are well documented and can be easily reproduced in virtual auditory display through binaural rendering [4], those allowing listeners to determine the distance of a sound source are less understood. Literature on distance perception of sound sources [5, 6] reports that humans significantly underestimate the distance of far sources and overestimate the distance of near sources. They report at least four auditory cues involved in the mechanisms of distance auditory estimation:

- In open space, *intensity* plays a major role with familiar sounds, it ideally decreases by 6 dB with doubling of distance between the source and listener. For unfamiliar sound sources, this cue is insufficient as it is confounded with the level of the sound itself [7].
- *Direct-to-reverberant energy ratio* is also an important cue in reflective and indoor environments. Mershon and King [8] have shown that distance perception is greater in reverberant environments compared to anechoic environment. Contrary to intensity, reverberation can allow the listener to make an absolute judgment of distance.
- If the listener has enough familiarity with the sound, the *spectrum* may convey distance cues as well. The spectral filtering, especially effective for far distances (particularly in the upper part of the auditory range) is induced by the absorption properties of the air and the eventual multiple reflections over non-ideal surfaces, which help one to estimate the distance of a sound source[9].
- For nearby sources, Brungart [10] has highlighted the importance of *binaural differences* in both intensity and time that are no longer independent of radial distance, as they are for far field planar waves. A study by Shinn-Cunningham *et al.* [11], provides a detailed analysis of binaural cue variations for nearby sound source location.

Despite the multiplicity of distance perception cues, the synthesis of range information in auditory display still remain a major issue and leads to poor quality results, especially for near field sound sources.

In an attempt to provide a linear relationship between perceived and physical distance, Devallez *et al.* [12] modeled a virtual listening environment consisting of a trapezoidal membrane with specific absorptive properties at the boundaries. This approach has been more recently extended by Fontana and Rocchesso [**?**] who studied the effect of exaggerating the acoustic cue of the reverberation by placing a real sound source in a pipe. They also demonstrated the possibility of creating flexible and virtual models for distance rendering with a simple physical system such as the acoustic pipe [13].

In the context of near field guidance (for distances inferior to 1.5 meters), distance perception is quite limited compared to the required precision. Instead of linearizing or exaggerating distance acoustics cues, this study aims to explore the influence of adding new acoustic cues for distance perception. It consists of representing distance cues instead of simulating them exactly. This can be realized through the use of sonification techniques.

In [14], Kramer defined sonification as "the use of non-speech audio to convey information or perceptual data". Many studies

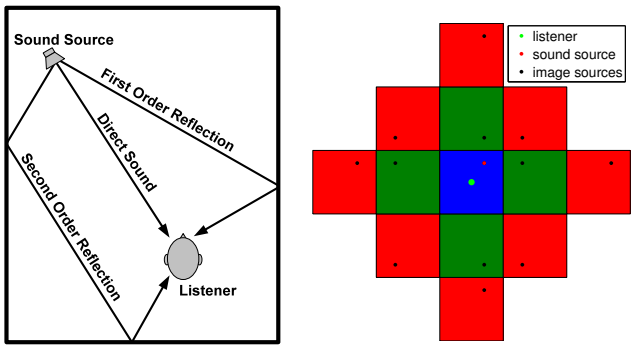
have investigated methods of conversion from data to sound. The Sonification Handbook [15] provides a good introduction to various methods. In this study, a parameter mapping sonification approach was used. This method consists in representing changes in data dimension through an auditory variation [14, 16]. Most existing parameter mapping sonification applications use pitch, time, loudness, or timbre as the principal mapping parameters applied on sound synthesis. While the transfer function between sonified data and sound synthesis parameters is very easy, one problem is that the sounds produced can be unpleasant and irritating for daily use.

In the past few years, despite the development of many sound interfaces, aesthetic and user acceptance issues have been absent from the scope of most research. Very few studies have investigated the customization of sound information by the user and its impact on the effectiveness and efficiency of the system. In [17], the authors worked on the aesthetics of sonification and found that musical sounds were more pleasant and appropriate than natural sounds. In [18], Brungart and Simpson describe the design of an audio display that modified the acoustic properties of an arbitrary audio input signal (e.g. pilot-selected music) to provide the pilot with information about the altitude of the aircraft.

In this article, the concept of parameter mapping sonification is extended to the use of any type of audio signal by mapping the parameters to audio effects (these are then applied to the sound). In this concept, the data no longer relies on the sound parameters but on the audio effect parameters. This allows for the application of the sonification metaphor to any type of sound while maintaining coherency with the data displayed. Applying this concept, three sound effect metaphors were created and initially evaluated with a near-field localization test designed with laboratory sounds.

2. SOUND EFFECT METAPHORS

In the context of a commercial project, several constraints are imposed on the development of the prototype and therefore on the distance sonification design. First of all, the use of binaural sound display imposes the use of large spectrum sound samples (to increase HRTF cues perception) with sharp attacks (to improve ITD perception). Then, the design of an accessible, aesthetically pleasing, and ergonomic device takes into account the end user's needs in terms of output user interface. These were evaluated using several questionnaires as well as a creativity session held with six visually impaired participants (see [1] for further details). In general, the visually impaired panel did not favor the use of sound as a method of guidance. In addition to the sound environment-masking problem due to the use of headphones, they reported a severe fatigue from the kind of sounds generally used (such as beeps, noise, and tones) in interfaces, and to the excessive length of messages in the case of text-to-speech based systems. As sound information may interfere with natural auditory cues in the real environment and cause supplementary cognitive load, the amount of information provided should be minimal, presenting only what is necessary and sufficient to aid the user. Presented messages should be highly efficient and minimally intrusive. The level of detail and display frequency of messages must be adjustable by the user. The sounds must be short and different from urban environmental sounds. One of the most important results of these investigations on user needs was the differing desires of system sounds amongst potential users. Some users asked for electronic sounds (such as video game sounds) in order to easily differentiate



(a) (b)

Figure 1: (a) Sound path in a room. (b) 2D schematization of the image-sources method. The simulated room is in blue, first order reflections are located in green areas and second order reflections are located in red areas. The listener is a green •, the source a red •.

them from the natural ambient sounds, while others preferred decontextualized natural sounds (animal, sea, cave, or forest sounds) or instrumental sounds. Regarding these results, it was not possible to find a general agreement on the types of sounds to use for the design of a navigation aid. Instead, a decision was made to design the sonification device using a customizable sound strategy.

2.1. Effect based sonification

To answer all of these constraints, distance sonification was designed as a digital audio effect applicable to the sound. With this concept, the distance is mapped to one or several parameters of the audio effect and the resulting sound pattern is thus distance dependent. This method allows for the design of several distance metaphors while leaving the user the possibility to customize the actual sounds of the interface. Furthermore, it has the advantage that once the metaphors are understood and learned, the user is able to change the sounds without relearning the sonification mapping.

On the basis of this idea, three distance metaphors were developed. The first one consists of reproducing a natural perceptual phenomena (sound reflection from walls), based on a simple room acoustic simulation. The other two metaphors are symbolic. There is no ecological link between the effect and the parameter represented. These metaphors are defined in the next section with the chosen mapping corresponding to the experimental setup, detailed in Sec. 3.

2.2. Early Reflection (ER)

As explained in Sec. 1, several studies highlighted the improvement of distance perception using reverberation cues [8, 20]. In [21], Begault showed the benefit of an artificial reverberation in a virtual auditory display. The addition of room reverberation led to better externalization and distance perception of the sound source, but slightly decreased azimuth localization performance. From literature on distance perception of nearby sources, a hypothesis was made that distance perception of sound sources in peripersonal space is improved by early reflections [22].

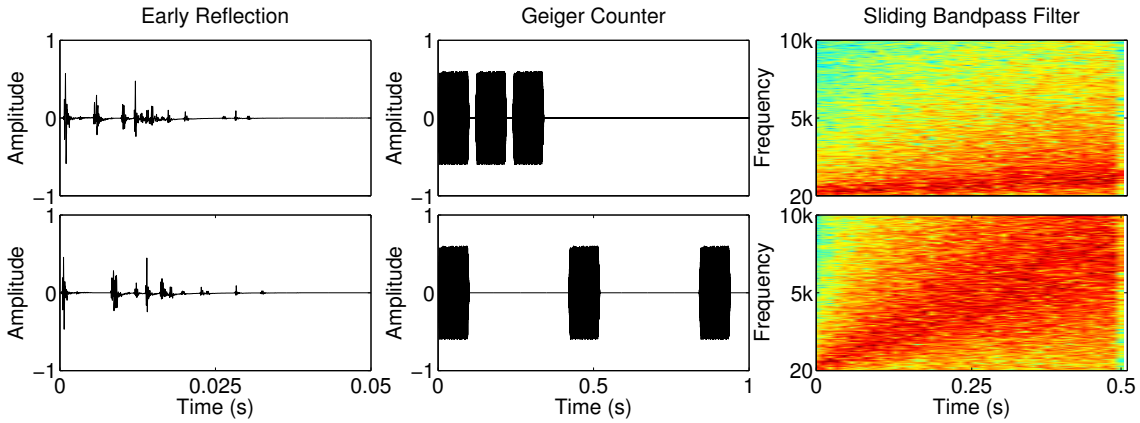


Figure 2: Representation of the sound resulting from the application of the three effect metaphors for two distances: $top = 0.6\ m$, $bottom = 1.5\ m$; (Left) Impulse response of the Early Reflection effect metaphor. (Center) Geiger Counter effect metaphor applied to a 10 ms burst. (Right) Spectrogram of the sounds resulting from the Sliding Bandpass Filter effect metaphor applied to a 0.5 sec burst.

The concept of this metaphor is therefore to create an effect based on the simulation of spatialized early reflection of second order (ie, reflecting off of one or two walls, considering an omnidirectional sound source, see Fig. 1) for a given room. In order to improve distance perception through the increase of natural audio cues with the simulation of room reverberation without decreasing the horizontal localization performances, a decision was made to simulate only early reflections. The image-source simulation method was used to simulate the early reflections [23]. Each reflection (called image-source) is a copy of the primary sound source coming from a different location. It is attenuated as a function of distance and filtered according to the absorption characteristics of the walls it encounters. These reflections allow for spatial information multiplication through the binaural spatialization of each reflection in addition to the direct sound source.

For the experiment, early reflections are based on the acoustic response of a $5 \times 5 \times 3m^3$ room. The head of the listener is placed at the center of this virtual room at a height of 1m40. 24 image-sources (6 first order reflections and 18 second order reflections) are necessary to simulate first and second order reflections. Their positions are calculated in real-time. Each source is filtered one or two times (depending on the number of walls encountered), then delayed according to the difference between their trajectory lengths and the trajectory of direct sound. In order to reduce computational time due to binaural rendering, the 24 sources are spatialized using a third order ambisonic method rendered over 12 virtual loudspeakers. These virtual loudspeakers surrounding the subject are then spatialized with binaural synthesis at classic positions on a sphere (for more details, see [24, 25]). The resulting binural signal is then mixed with the binauralized direct sound signal. Fig. 2 (left) represents the impulse response of this metaphor effect for two different distances (0.6 m and 1.5 m).

2.3. Geiger Counter (GC)

One of the first sonification applications was the Geiger counter, invented by Hans Geiger in the early 1900's. It consists of increasing the rate of a generated "beep" in proportion to the intensity of non-visible radiation. This well-known metaphor has been successfully tested in a number of sonification applications, and has now become a part of everyday life, used for several commercial applications. For example, it is used on some vehicle reversing/parking aids, which are intended to avoid collisions when reversing a vehicle. As an obstacle comes closer, the warnings become more strident and insistent.

To increase the perception of distance, this effect consists of repeating the stimulus three times and varying the time interval between each repetition as a function of distance. Thus, the closer the target is, the faster the repetition.

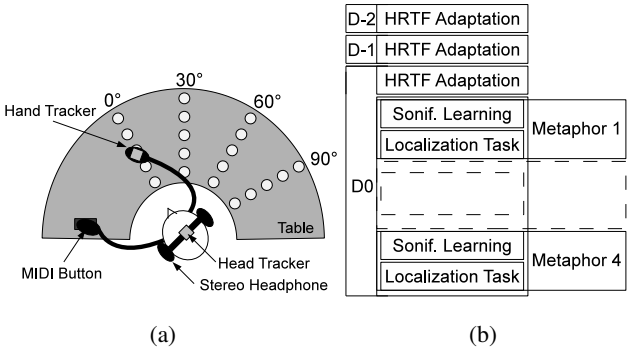
This mapping was chosen so as to avoid any overlap of sounds when the target is near the user, thus the variations were sufficiently noticeable. Time repetitions are therefore of 20 ms at 0.6 m and of 320 ms at 1.5 m, the evolution between these two distances is linear. The sound signal resulting from the application of this metaphor to a 10 msec burst for two different distances (0.6 m and 1.5 m) is presented in Fig. 2 (center).

2.4. Sliding Bandpass Filter (SBF)

Several studies have shown that the used of pitch in data sonification was easily understandable and efficient [26]. The idea of this metaphor is to transpose this sonification concept to an audio effect applicable to any type of sound.

This effect is created using a band-pass filter with a time sliding central frequency and a time varying bandwidth, such that so the quality factor $Q = \Delta f / f$ remains constant (where Δf is the bandwidth and f the central frequency). The initial central frequency of the filter (at T=0 sec, beginning of the sound) is fixed to 200 Hz regardless to the distance. The final central frequency of the filter (at T= sound length, end of the sound) increase proportionally with distance. With this effect, a noise burst will sound as a noisy chirp with a higher final frequency depending on the distance.

For the experiment, the quality factor was fixed to $Q = \Delta f / f = 2$, the final frequency was fixed to 1 kHz for a target placed at 0.6 m and to 8 kHz for a target at 1.5 m. The evolution of the final frequency according to the variation of the distance is linear. Fig. 2 (right) represents the spectrogram of the sound resulting from this effect applied to a white noise burst of 0.5 sec for two different distances (0.6 m and 1.5 m).



(a) (b)

Figure 3: (a) Experimental setup. Small circles = sound source positions (b) Timeline of the experiment.

3. METHODS

3.1. Participants

A total of 16 adult subjects not visually impaired (3 women and 13 men, mean age 28 ± 6) served as paid volunteers; An audiogram was performed on each subject before the experiment to ensure that their audition was normal (> 15 dB(HL)). All were naive regarding the purpose of the experiment and the sets of spatial positions selected for the experiment.

3.2. Apparatus

A diagram of the setup for the experiment is shown in Fig. 3, with a timeline of the experimental procedure. The first three stages consist of three adaptation sessions with the non-individual HRTFs, see sec 3.3. The next stages consist of the evaluation of each sonification condition with a localization task. During the localization sessions, subjects were seated on a swivel chair located at the center of a wooden circular table of 90 cm in diameter.

The subjects were equipped with a stereo open ear headphone (model Sennheiser HD570) tracked with a 6-DoF position/orientation magnetic sensor positioned on the top of the headphone. They held a position sensor in their dominant hand and interacted with the system using a MIDI button with their other hand. The position of the hand was calculated relative to the tracked center of the head. No headphone equalization was used.

The stimulus used was rendered via a set of non-individual HRTF measured on a KEMAR mannequin (describe in sec 3.3). It was brief to avoid head movement effects and consisted of a train of three, 40 ms Gaussian broadband noise bursts (50 – 20000 Hz) with 2 ms Hamming ramps at onset and offset and 30 ms of silence between each burst. This stimulus was chosen following Dramas *et al.* [27] where the effect of repetition and duration of the burst on localization accuracy was analyzed. Their results showed an improvement of the accuracy between three repeated 40 ms bursts and a single 200 ms burst. The overall level of the train was approximately 60 dBA measured at the ears for a binaural sound source rendered at 50 cm in front of the subject (0° in azimuth and 0° in elevation).

3.3. KEMAR HRTF

The HRTF of a KEMAR mannequin was measured at IRCAM's anechoic chamber. In order to render all the localization test's positions, it was necessary to measure the HRTF over the entire sphere. The set used contained measures from -90° to 90° in elevation in steps of 5°, and from -180° to 180° in azimuth in steps of 15°. These measures are more precise in elevation, otherwise they have the same characteristics as HRTFs of the LISTEN database [28].

In order to improve the localization performances of the subject with the binaural rendering using this non-individual HRTF, three adaptation sessions of 12 min were conducted according to the method proposed by Parseihian and Katz [29]. Briefly, this method consists of a training game allowing the subject to do a quick exploration of the spatial map of the virtual rendering by an auditory-kinesthetic process. These training sessions were performed three days in a row, twelve minutes per day, the last session being immediately followed by the main experiment.

3.4. Procedure

The experiment was divided into four blocks of 80 trials, each block lasting approximately 15 min. Each block corresponds to a different distance metaphor condition. In order to evaluate the improvement effect of each sonification metaphor, a block of trials without sonification (i.e. only binaural rendering) served as a reference for localization performance. The four blocks are called: *control* (for no sonification), *geiger counter (GC)*, *sliding band-pass filter (SBF)*, and *early reflection (ER)*. For each subject, the blocks were presented in a random order so as to counterbalance any potential task learning effect. Each block of trials began with a short learning session of the sonification metaphor during which the sound was repeated every two seconds. The aim of this learning session was to accustom the subjects to the distance metaphor by allowing them to interact with the distance with an auditory-kinesthetic process. First, for the subject to be aware of the distance ranges and the variations of the acoustic cues, he was asked to move his hand from the inside to the outside of the table and then return, thus two times for two different directions (frontal and lateral). Then, for a periode of one minute, the subject had total control of a virtual sound source spatialized at his hand position and was asked to freely explore the entire surface of the table.

The localization task consisted of reporting the perceived position of a static spatialized sound sample using a hand placing technique validated by a MIDI button. Each subject was instructed to orient himself straight ahead and to keep his head fixed, in a reference position at the center of the system, 0.65 m over the table, during the brief sound stimulus presentation. Before each trial, the subject's head position was automatically compared to the reference position and the subject was asked to correct his position if there was no concordance(± 5 cm for the position and $\pm 3^\circ$ for the orientation). After presentation of the stimulus, each subject was instructed to place his hand on the table at the current position of the perceived sound source location and to validate the response with the MIDI button. The subjects were placed in the system in order to use their dominant hand. The perceived position was calculated between the initial head position/orientation when the stimulus was played and the final hand position when the listener validated the target. No feedback was given to the subject regarding the actual target position.

Condition	Control	ER	GC	SBF
Regression slope	0.14 (.09)	0.06 (.13)	0.64 (.29)	0.50 (.20)
Goodness-of-fit	0.66 (.26)	0.27 (.31)	0.96 (.05)	0.92 (.12)

Table 1: Mean linear regression analysis and goodness-of-fit criteria r^2 of the perceived distance. Variances shown in parentheses.

A total of 20 positions (5 different distances relative to the head: 0.73 m, 0.80 m, 0.88 m, 0.97 m, and 1.07 m and 4 azimuths: 0°, 30°, 60°, and 90°, see Fig. 3), were randomly presented with 4 repetitions each. Subjects had to localize a total of 80 targets and were naive with respect to the set of spatial positions selected for the experiment.

4. RESULTS

The contribution of the sonification metaphors on the perceived distance was analyzed by comparing the distance and azimuth errors of each metaphor (*geiger counter*, *sliding bandpass filter*, and *early reflection*) to those of the control reference condition without sonification (*control*). Because of validation problems with some participants, all trials with a hand position outside the table have been removed from the analysis. Some front/back confusion errors were noticed for rendered sources at 30° and 60°. Since this paper is focused on distance perception, these confusions were corrected before data analysis.

4.1. Effect of the metaphors on the perceived distance

Fig 4 shows the average mean response of perceived source distance as a function of virtual source distance and the mean of linear regression for each condition. It highlights a tendency to overestimate sound distance for the two nearest rendered distances and to overestimate it for the others. It can also be noted that results for *control* and *early reflection* were poorer than those for *GC* and *SBF* conditions. A linear regression analysis was performed on these results. The mean and standard deviation across subjects of the slope of the regression line and goodness-of-fit criteria r^2 for each condition are shown in Table 1. Regression slope lines were far from the unity expected for a perfect distance perception of virtual sound for the *control* and *ER* conditions. For these two conditions there was no real perception of distance. The results for the *SBF* and the *GC* conditions were better with regression slopes nearer to unity but with larger inter-subject variability (highlighted by the large standard deviation).

These results are confirmed by the boxplot of relative distance error shown in Fig. 5. Indeed, the mean errors of the *GC* and the *SBF* conditions are approximately 5 cm lower than those of the *control* and the *ER* conditions. A repeated measures ANOVA was performed on the mean distance error, taking into account three within-subjects factors: metaphor condition (4 levels, fixed factor), rendered distance (5 levels, fixed factor) and rendered azimuth (4 levels, fixed factor). It showed a significant effect of the metaphor condition ($F(3,42) = 19.76, p < 0.001$), the rendered distance ($F(4,56) = 12.01, p < 0.001$) and the rendered azimuth ($F(3,42) = 9.32, p < 0.001$). A Duncan test on categories showed significant differences between *control* and *GC* conditions ($p = 6.10^{-5}$) and between *control* and *SBF* conditions ($p = 2.10^{-4}$). The comparison of *control* and *ER* conditions showed no-significant effects ($p = 0.59$). For the rendered positions, a Duncan test on distance revealed significant differences between the farther distance and the others (highlighting poorer performances for farther distances), and a Duncan test on azimuth revealed significant differences between the lateral angle 90° and the others (highlighting better performance for lateral positions).

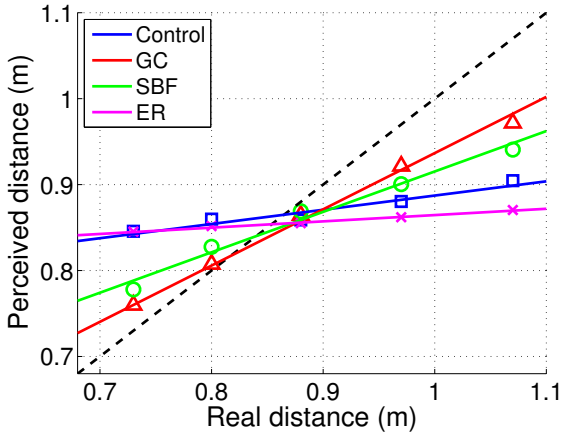


Figure 4: Perceived distances as a function of rendered distance for each sonification condition. «□, △, ○, ×»: Mean under each condition. Lines: Mean of linear regression.

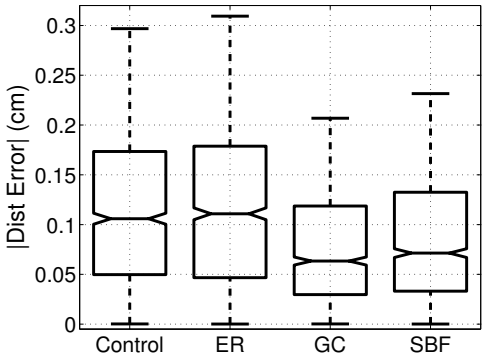


Figure 5: Boxplot of the relative distance error for each metaphor.

Angle	0°	30°	60°	90°
Control	0.13 (.10)	0.11 (.09)	0.11 (.08)	0.09 (.08)
ER	0.11 (.10)	0.10 (.09)	0.12 (.09)	0.10 (.08)
GC	0.07 (.08)	0.06 (.07)	0.06 (.07)	0.06 (.06)
SBF	0.09 (.09)	0.08 (.08)	0.07 (.07)	0.06 (.06)

Table 2: Mean distance error (in m) per angle and metaphor. Variances shown in parentheses.

A thorough study of the perceived distance error while taking into account the effect of the rendered azimuth is shown for all conditions together in Fig. 6 and for each condition in the Table 2. The boxplot highlights better performance for distance perception for lateral sound sources than for frontal sound sources. Regarding Table 2, this slight improvement in performance for lateral sound sources almost appeared for the *control* and the *SBF* conditions

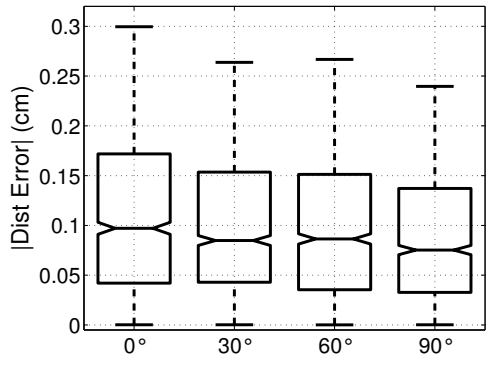


Figure 6: Boxplot of the relative distance for all conditions as a function of azimuth angle.

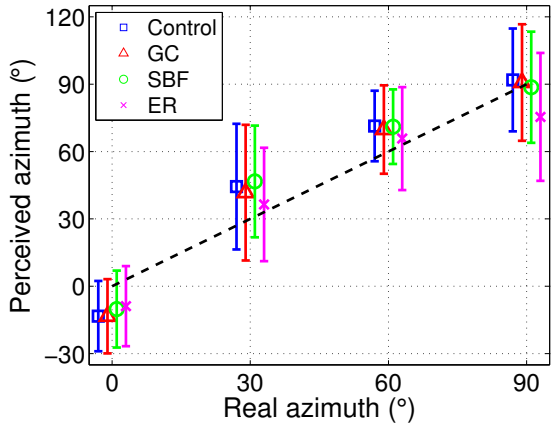


Figure 7: (a) Perceived azimuth as a function of rendered azimuth for each sonification condition. «□, △, ○, ×»: Mean for each condition. Vertical lines: Standard deviation for each modality. For the sake of readability, results corresponding to the different conditions have been slightly horizontally shifted.

(but with a large standard deviation).

4.2. Effect of the metaphors on the perceived azimuth

Although this was not the primary aim of this study, it is interesting to look at the effect of the sonification metaphors on the perceived azimuth angles. Fig. 7 shows the average mean response of perceived source azimuth as a function of virtual source azimuth for each condition. It highlights a large standard deviation mainly at 30° and 90°, and a shift of 10° for frontal sources. Regarding each condition, it appears that the metaphors did not affect the azimuth performances except for lateral sound sources with the *ER* condition.

The mean azimuth error was $20 \pm 15^\circ$. Performing a repeated measurement ANOVA on the relative azimuth error for each metaphor, mixing all the positions, showed no significant effect on the metaphor condition ($F(3, 45) = 0.206, p = 0.89$).

5. DISCUSSION

Regarding the results, of the three designed metaphors, only two most were effective (the *geiger counter* and the *sliding bandpass filter* metaphors) than the control condition of pure binaural anechoic synthesis. Compared to the control condition without sonification (condition whose performances were almost zero for the rendered distances of the experiment), these two effect metaphors improved distance perception significantly. The superiority of the *geiger counter* metaphor over the *sliding bandpass filter* could be explained by their mapping parameters. Indeed, the mapping of the *sliding bandpass filter* metaphor was linear, whereas our perception of frequency is logarithmic. It seems that the variation range of the frequency was not wide enough to be sufficient for a complete rendering of the distances.

Contrary to what was expected, the *early reflection* metaphor failed to improve the distance perception and led to poorer performances than the *control* condition. Furthermore, directional localization at 90° was degraded by this metaphor, which was not the case with the other conditions. To explain this, several observations can be made. First, the chosen model with only early reflections of the first and the second order was too simple, and the absence of the reverberation tail may have affected perception by creating an abnormal situation. Second, all of the studies reporting an improvement of the perceived distance with early reflections were conducted with distances superior to one meter. These cues are perhaps not effective for the shorter distances used in this study.

For all the conditions, but mainly in *control* and *sliding bandpass filter* conditions, perceived distance performance was better for lateral sound sources (especially at 90° azimuth). This improvement, appearing in all conditions, seems to be specific to the binaural rendering. Indeed, in this experiment, distance was linked to elevation as the subjects were 0.65 m over the table. This results in an elevation of -37° for the longest distance and of -63° nearest source. For these elevations, the influence of the torso is more important for lateral sources than frontal sound sources. This probably influenced distance perception. These results are confirmed by the results of a study by Kopco and Shinn-Cunningham [30] that showed better performance for distance perception for lateral sound sources using real sound sources. This result is mainly explained by the variation of Interaural Level Difference (ILD) as a function of distance for lateral sources (due to the shadowing effects of the head) and by the absence of variation for frontal sources (since the ILD is equal to zero).

Regarding the results for directional localization, except for the condition *early reflection* at 90°, there was no effect of distance metaphor on the perceived azimuth. The directional errors were slightly poorer than results with real sound sources (for distance between 0.5 and 1 m, and elevation below -20°, Brungart *et al.* [10] obtained a mean azimuth error of 11°). With an average error of 20°, these performances are not so bad considering that the HRTF set used in the experiment contained azimuthal measures at 15° intervals, as well as being non-individualized.

Since the setup of this experiment differs from how previous studies have been organized, precise comparison is impossible. For localization of real nearby sound sources in anechoic environments, distance performance obtained by Brungart *et al.* [10] were from a regression slope of 0.3 for frontal sources to 0.8 for lateral sources. While simulating nearby sound sources with binaural room impulse responses recorded in a reverberant environment, Kopco and Shinn-Cunningham [30] obtained better perfor-

mance with a mean of regression slope of 0.6 for frontal and 0.8 for lateral distance perception. In this study, with only binaural conditions there was no real perception of distance (regression slope of 0.14). This can be explained by the used HRTFs that were non-individualized and were actually measured at a distance of two meters, so they do not naturally contain near field binaural cues despite attempts to improve performance. In addition, the source positions used in this study, all being in the lower hemispphere, may bias results due to the potential difficulty in this region. With the *geiger counter* and the *sliding bandpass filter* metaphors (regression slope of 0.64 and 0.5), the results approach the performances obtained in [10], thereby highlighting the effectiveness of the adopted method for the sonification.

6. CONCLUSION

The aim of this study was to design and evaluate several metaphors of sound source distance sonification for virtual auditory display. In order to respond to user needs, the designed sonification needed to be independent of the actual sound as well as easy to learn. On the basis of these constraints, the concept of sound effect based sonification was introduced. This new sonification concept consists of the application of an audio effect, whose parameters are dependent on the data to sonify, to any type of sound. With this method, the information is contained in the audio effect and not in the sound. On this basis, three distance metaphors were created and evaluated with sound localization experiments. These experiments underline the contribution of these metaphors to distance perception compared to a control reference condition consisting solely of anechoic binaural rendering. The results highlight a significant improvement of the distance perception with two of the tested metaphors (the *geiger counter* and the *sliding bandpass filter*) in spite of only a short learning period (one minute). It would be interesting to explore the mapping of these metaphors in more detail and their effects on users performance.

The success of these two effect metaphors in improving near field distance perception shows the equivalence of the effect metaphor concept to the traditional parameter mapping sonification applied to sound synthesis. This is a positive result regarding user acceptance of the sonification, which often suffers from a lack of aesthetics.

Since this study was focused on the efficiency of the effect metaphors with "laboratory sounds" (noise burst), further experiments should now be carried out to validate their efficiency with "real sounds" (ecological, instrumental, or electronic sound) in order to approach the real situations and determine if it meets users requirements. Through further studies, it will be interesting to modify traditional parameter mapping sonification strategies into effect mapping sonifications. This will allow for expanded testing based on the findings of this emerging research field.

7. ACKNOWLEDGMENT

This work was supported in part by the French National Research Agency (ANR) through the TecSan program (project NAVIG ANR-08TECS-011) and the Midi-Pyrénées region through the APRRTT program. The authors would like to thank all of the subjects for participating in the experiment, and IRCAM for use of their measurement facilities regarding the acquisition of the KEMAR HRTF. Finally, thanks to Omnihead (http://www.univ-brest.fr/mstis/omnihead/accueil.htm) for the use of the KEMAR dummy head.

8. REFERENCES

[1] B. Katz, F. Dramas, G. Parseihian, O. Gutierrez, S. Kammoun, A. Brilhault, L. Brunet, M. Gallay, B. Oriola, A. Auvray, P. Truillet, M. Denis, S. Thorpe, and C. Jouffrais, "Navig: Guidance system for the visually impaired using virtual augmented reality," *Journal of Technology and Disability*, vol. 24, 2012 (in press).

[2] F. Dramas, S. Thorpe, and C. Jouffrais, "Artificial vision for the blind: A bio-inspired algorithm for objects and obstacles detection," *International Journal of Image and Graphics*, vol. 10, no. 4, pp. 531–544, nov 2010.

[3] B. Katz, E. Rio, and L. Picinali, "LIMSI Spatialization Engine," Inter Deposit Digital Number: F.001.340014.000.S.P.2010.000.31235.

[4] D. Begault, *3-D Sound for Virtual Reality and Multimedia*. Cambridge: Academic Press, 1994.

[5] J. Loomis, R. Klatzky, and R. Golledge, "Auditory distance perception in real, virtual and mixed environments," *Mixed Reality: Merging Real And Virtual Worlds*, 1999.

[6] P. Zahorik, D. Brungart, and A. Bronkhorst, "Auditory distance perception in humans : A summary of past and present research," *Acta Acoustica United with Acoustica*, vol. 91, no. February 2003, pp. 409 – 420, 2005.

[7] P. Coleman, "Failure to localize the source distance of an unfamiliar sound," *J. Acoust. Soc. Am.*, vol. 34, pp. 345–346, 1962.

[8] D. Mershon and L. King, "Intensity and reverberation as factors in the auditory perception of egocentric distance," *Attention, Perception, & Psychophysics*, vol. 18, no. 6, pp. 409–415, 1975.

[9] J. Blauert, *Spatial Hearing*. Cambridge: MIT Press, 1996.

[10] D. Brungart, N. Durlach, and W. Rabinowitz, "Auditory localization of nearby sources. II. localization of a broadband source," *J. Acoust. Soc. Am.*, vol. 106, no. 4, pp. 1956–1968, 1999.

[11] B. G. Shinn-Cunningham, S. Santarelli, and N. Kopco, "Tori of confusion: Binaural localization cues for sources within reach of a listener," *J. Acoust. Soc. Am.*, vol. 107, no. 3, pp. 1627–1636, 2000.

[12] D. Devallez, F. Fontana, and D. Rocchesso, "Linearizing auditory distance estimates by means of virtual acoustics," *Acta Acustica united with Acustica*, vol. 94, no. 6, pp. 813–824, Sept 2008.

[13] F. Fontana, D. Rocchesso, and L. Ottaviani, "A structural approach to distance rendering in personal auditory displays," in *IEEE International Conference on Multimodal Interfaces (ICMI 2002)*, 2002.

[14] G. Kramer, *Auditory Display: Sonification, Audification and Auditory Interfaces*. Perseus Publishing, 1993.

[15] T. Hermann, A. Hunt, and J. Neuhoff, Eds., *The Sonification Handbook*. Berlin, Germany: Logos Publishing House, 2011. [Online]. Available: http://sonification.de/handbook

[16] B. N. Walker and G. Kramer, "Mappings and metaphors in auditory displays: An experimental assessment," in *Proceedings of the 3rd International Conference on Auditory Display (ICAD96)*, S. P. Frysinger and G. Kramer, Eds., 1996.

[17] C. Sikora, L. Roberts, and L. Murray, "Musical vs. real world feedback signals," in *Conference companion on Human factors in computing systems*, ser. CHI '95. New York, NY, USA: ACM, 1995, pp. 220–221.

[18] D. Brungart and B. Simpson, "Design, validation, and in-flight evaluation of an auditory attitude indicator based on pilot-selected music," in *Proceedings of the International Conference on Auditory Display (ICAD2008)*, Paris, France, 2008.

[19] L. Brunet, "Étude des besoins et des stratégies des personnes non-voyantes lors de la navigation pour la conception d'un dispositif d'aide performant et accepté (Needs and strategy study of blind people during navigation for the design of a functional and accepted aid device)," Master's thesis, Departement of Ergonomics, Universitée Paris-Sud, Orsay, France, 2010.

[20] S. Nielsen, "Auditory distance perception in different rooms," in *Audio Engineering Society Convention 92*, March 1992.

[21] D. Begault, "Perceptual effects of synthetic reverberation on three-dimensional audio systems," *J. Audio Eng. Soc*, vol. 40, no. 11, pp. 895–904, 1992.

[22] G. Kearney, M. Gorzel, H. Rice, and F. Boland, "Distance perception in interactive virtual acoustic environments using first and higher order ambisonic sound fields," *Acta Acustica united with Acustica*, vol. 98, no. 1, pp. 61–71, 2012.

[23] J. Allen and D. Berkley, "Image method for efficiently simulating small-room acoustics," *Acoustical Society of America Journal*, vol. 65, pp. 943–950, Apr. 1979.

[24] A. McKeag and D. S. McGrath, "Sound field format to binaural decoder with head tracking," in *Audio Engineering Society Convention 6r*, 1996.

[25] M. Noisternig, A. Sontacchi, T. Musil, and R. Holdrich, "A 3d ambisonic based binaural sound reproduction system," in *Audio Engineering Society Conference: 24th International Conference: Multichannel Audio, The New Reality*, 2003.

[26] L. M. Brown, S. Brewster, R. Ramloll, M. Burton, and B. Riedel, "Design guidelines for audio representation of graphs and tables," in *Proceedings of the 9th International Conference on Auditory Display (ICAD2003)*. Boston, USA: Boston University Publications Production Department, 2003, pp. 284–287.

[27] F. Dramas, B. F. Katz, and C. Jouffrais, "Auditory-guided reaching movements in the peripersonal frontal space (poster)," in *Acoustics 08, Paris, 29/06/2008-04/07/2008*, vol. 123. J. Acoust. Soc. Am., 2008, p. 3723.

[28] "IRCAM LISTEN HRTF database," http://recherche.ircam.fr/equipes/salles/listen/.

[29] G. Parseihian and B. Katz, "Rapid head-related transfert function adaptation using a virtual auditory environment," *J. Acoust. Soc. Am.*, vol. 131, no. 4, 2012.

[30] N. Kopco and B. Shinn-Cunningham, "Effect of stimulus spectrum on distance perception for nearby sources," *J. Acoust. Soc. Am.*, vol. 130, no. 3, pp. 1530–1541, June 2011.

SOUND FOR 3D CINEMA AND THE SENSE OF PRESENCE

Cédric R. André,
Jean-Jacques Embrechts,
and Jacques G. Verly

INTELSIG Laboratory
University of Liège, Liège, Belgium
C.Andre@ulg.ac.be

Marc Rébillat

Laboratoire Psychologie de la Perception, CNRS,
Université Paris Descartes, Paris, France
and Département d'Études Cognitives,
École Normale Supérieure, Paris, France.
marc.rebillat@ens.fr

Brian F.G. Katz

LIMSI-CNRS
Orsay, France
brian.katz@limsi.fr

ABSTRACT

While 3D cinema is becoming more and more established, little effort has focused on the general problem of producing a 3D sound scene spatially coherent with the visual content of a stereoscopic-3D (s-3D) movie. As 3D cinema aims at providing the spectator with a strong impression of being part of the movie (sense of presence), the perceptual relevance of such spatial audiovisual coherence is of significant interest. Previous research has shown that the addition of stereoscopic information to a movie increases the sense of presence reported by the spectator. In this paper, a coherent spatial sound rendering is added to an s-3D movie and its impact on the reported sense of presence is investigated. A short clip of an existing movie is presented with three different soundtracks. These soundtracks differ by their spatial rendering quality, from stereo (low spatial coherence) to Wave Field Synthesis (WFS, high spatial coherence). The original stereo version serves as a reference. Results show that the sound condition does not impact on the sense of presence of all participants. However, participants can be classified according to three different levels of presence sensitivity with the sound condition impacting only on the highest level (12 out of 33 participants). Within this group, the spatially coherent soundtrack provides a lower reported sense of presence than the other custom soundtrack. The analysis of the participants' heart rate variability (HRV) shows that the frequency-domain parameters correlate to the reported presence scores.

1. INTRODUCTION

Although many movies are now produced in stereoscopic 3D (s-3D), the sound in these movies is still most often mixed in 5.1 surround. The information conveyed in this format is rarely accurately localized in space. The dialogs, for example, are confined to the front center channel [1]. Therefore, the sound mix does not provide the moviegoer with a 3D sound scene spatially consistent with the visual content of the s-3D movie.

As 3D cinema aims at providing the spectator with a strong impression of being part of the movie, there is a growing interest in the sense of presence induced by the media. Presence (or more accurately, telepresence) is a phenomenon in which spectators experience a sense of connection with real or fictional environments and with the objects and people in them [2]. Previous research has shown that the addition of stereoscopic information to a movie increases the sense of presence reported by the spectators [3]. It is hypothesized that the spatial sound rendering quality of an s-3D movie impacts on the sense of presence as well.

This study considers, in the cinema context, the cognitive differences between a traditional sound rendering (stereo), and a highly precise spatial sound rendering (Wave Field Synthesis or WFS). In particular, it will be examined whether a higher spatial coherence between sound and image leads to an increased sense of presence for the audience. The current study therefore presents the results of a perceptual study using a common video track and three different audio tracks. Using a post-stimuli questionnaire based on previous reports regarding the sense of presence, various cognitive effects are extracted and compared.

2. THE SMART-I^2

The present study was carried out using an existing system for virtual reality called the SMART-I^2 [4], which combines s-3D video with spatial audio rendering based on WFS.

The SMART-I^2 system (Fig. 1) is a high quality 3D audiovisual interactive rendering system developed at the LIMSI-CNRS in collaboration with *sonic emotion*[1]. The 3D audio and video technologies are brought together using two Large Multi-Actuator Panels, or LaMAPs ($2.6\ \mathrm{m} \times 2\ \mathrm{m}$), forming a "corner" that acts both as a pair of orthogonal projection screens, and as a 24 channel loudspeaker array. The s-3D video is presented to the user using passive stereoscopy, and actuators attached to the back of each LaMAP allow for a WFS reproduction [5] in a horizontal window corresponding to the s-3D video window. WFS [6] is a physically based sound rendering method that creates a coherent spatial perception of sound over a large listening area by spatially synthesizing the acoustic sound field that real sound sources would have produced at chosen locations [4]. The $20\ \mathrm{cm}$ spacing between the actuators corresponds to a spatial aliasing frequency of about $1.5\ \mathrm{kHz}$, the upper frequency limit for a physically correct wavefront synthesis, accounting for the loudspeaker array size and the extension of the listening area [7]. It is not a full 3D audio system, since, due to the use of a linear WFS array, the rendering is limited to the horizontal plane. Azimuth and distance localizations accuracies of sound events in the SMART-I^2 were previously verified by perceptual experiments and are globally consistent with real life localization accuracy [4].

There is a distinction in the SMART-I^2 between the direct and the reverberant parts of the sound. The direct sound is sent to the WFS rendering engine, which controls the actuators on the LaMAPs, while a Max/MSP based spatial audio processor (the Spat$\sim$, [8]) generates the reverberant portion, which is then fed to six surround loudspeakers and a subwoofer (Fig. 1).

In the SMART-I^2 system, the Spat$\sim$ is used to generate the reverberant field (processing load configuration `1a 8c 6r 0`, see [9] for more details). Each of the 16 input channels to the SMART-I^2 goes through the following DSP chain. The pre-processing

[1] www.sonicemotion.com

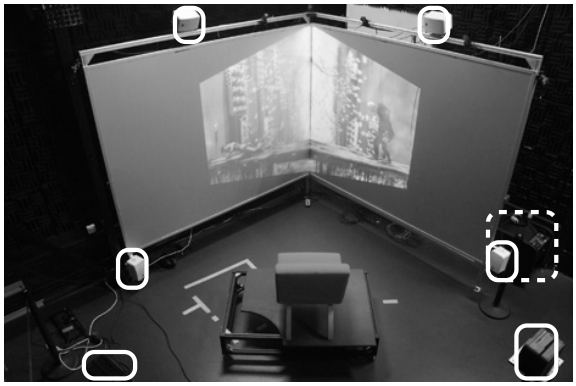


Figure 1: Photo of the SMART-I^2 installation for cinema projection highlighting the image correction for perceived plane projection at the experimental viewing position. —: Surround speakers, - -: Subwoofer.

of each source signal (Source∼) allows for the air absorption and the distance attenuation computations. The room simulator (Room∼) uses 8 internal feedback channels per source and uses a temporal division of the room response in early reflections, reflection clusters, and diffuse late reverberation. The directional encoding and distribution module (Pan∼) computes a pairwise intensity panpot for reproduction over a 6-loudspeaker horizontal array. The reverberation was adjusted to have an early decay time of 1.1 s, a reverberation time of 2.0 s, and a direct-to-reverberant ratio of -24 dB. This room response was not modified over the course of the movie.

The SMART-I^2 is currently capable of rendering in real-time 16 concurrent audio streams (sources), in addition to the Spat∼ room effect channels. The spatial position of these streams can be dynamically changed. In this study, the audio streams and their spatial positions were controlled using a sequence table, which identified the current audio files and their associated spatial coordinates.

The image was projected onto the corner of the SMART-I^2 to avoid any dissymetry in sound reproduction due to reflections coming from one side only. Since the goal was to approximate cinema conditions, it was necessary to compensate for this geometry, so that the projected 2D images appeared rectangular and planar from the subjects' viewpoint. The open source s-3D movie player Bino, which is compatible with the Equalizer library [10], was used to read the video stream. This allowed for projection onto the particular screen configuration, obtaining a result close to one that would be obtained on a regular planar screen, for a specifically defined viewing position. The difference was mainly seen at the top and bottom of the image, where trapezoidal or keystone distortion was visible when away from the experimental position (Fig. 1).

Due to cinema image aspect ratio, the projected image did not fill the whole surface of the two panels. Hence, the audio engine was capable of rendering objects which were effectively outside the video window. For example, for a spectator seated 3 m from the SMART-I^2 corner (Fig. 1), the horizontal field of view was about $61°$, and the audio field was about $119°$.

3. THE SELECTED MOVIE

It was decided to use an animation s-3D movie to carry out this study, rather than a real-image s-3D film. The reason was that the use of an animation movie allows for the automatic recovery of the exact spatial information of all objects present in the scenes from the source files.

The film selected for this project was "Elephants Dream"[2], an open movie, made entirely with Blender, a free open source 3D content creation suite[3]. All production files necessary to render the movie video are freely available on the Internet. For this pilot study, only the first three scenes of the movie were generated ($t = 00$ min 00 s to $t = 02$ min 30 s).

The first scene ($t = 00$ min 00 s) starts with the opening credits, where the camera travels upward until it reaches the first character's reflection in water. In the second scene ($t = 00$ min 27 s), the two characters are attacked by flying cables and there is a dialog. The third scene ($t = 01$ min 10 s) consists of the two characters running through a large room, being chased by mechanical birds .

Source position density plots were calculated for the three scenes (Fig. 2), indicating the positions where sources are present. It can be observed that most sources were frontal and centered, located just behind the screen plane. The second scene exhibits many lateral sources. In general, few sources are found in front of the screen, with only the third scene exploiting depth variations. The paths of the cables in the second scene and the birds in the third scene are the farthest positioned sources.

Contrary to the image source code, the audio track of the movie was only available in the final downmix version (stereo and 5.1), with some additional rough mixes of most of the dialogs and music tracks. The original multitrack audio master was not available. It was therefore necessary to create a new audio master with each object corresponding to a separate track, in order to allow for position coherent rendering. The aim was to recreate an audio track similar to the original track. The available dialog and music dry tracks were retained. The rest of the audio elements were created from libraries and recorded sounds, with one audio file per object.

The result was an object oriented multitrack audio master that contained individual audio tracks for each individual audio object, allowing for individual rendering positions to be defined and controlled. Details on the creation of the object-oriented audio and control tracks can be found in [11].

4. SOUNDTRACKS

4.1. Different spatial sound renderings

Three different soundtracks were used in this experiment. The first soundtrack is the original stereo soundtrack, termed `ST`. This soundtrack was rendered on the WFS system by creating two virtual point sources at $\pm 30°$ in the (virtual) screen plane, roughly at the left/right edges of the image. The object-oriented soundtrack, termed `WFS`, was the spatially coherent rendering. This new audiotrack was created specifically as part of this study, but was inspired by the original `ST` audiotrack. Due to the content differences between `ST` and `WFS`, an ideal stereo mix was constructed using the same metadata as in the `WFS` version. The panning of each object in this mix was automatically determined according to a sine panning law relative to the object's actual position (the same as the `WFS` version), and a corresponding r^{-2} distance attenuation factor was applied. This hybrid soundtrack, termed `HYB`, thus had the same content as the `WFS` track, but was limited in its spatial rendering quality. The `HYB` track was rendered over the same two virtual point sources as the `ST` track.

Due to differences between the soundtracks, a global equalization across the entire movie was inappropriate, and resulted in distinctly different perceived levels. Therefore, it was decided to equalize for an element that was common to all conditions. One character line,

[2] `www.elephantsdream.org`
[3] `www.blender.org`

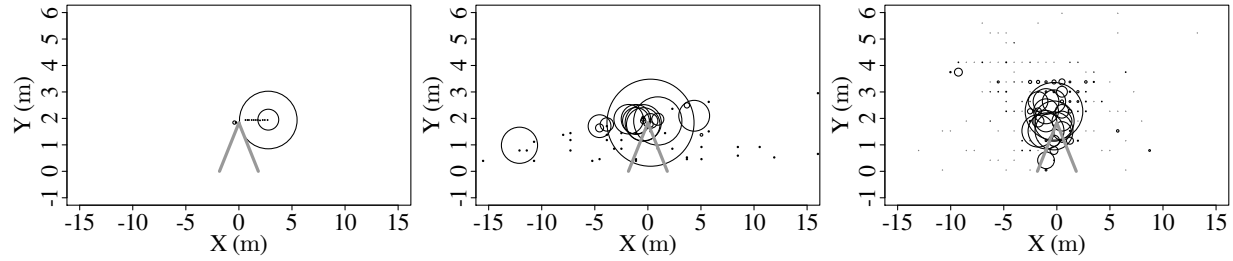


Figure 2: Bubbleplots of the sound source positions in the horizontal plane, taken every 5 frames for the first (left), second (center), and third (right) scenes. Diameters are proportional to the density of sources at that position. Some very distant source positions are not shown for clarity. Note: horizontal and vertical axes have different scales. The panels of the SMART-I^2 are represented by the inverted "V" which represent a 90° angle.

at $t = 00$ min 22 s, duration 4 s, was chosen as it was common to all three soundtracks (dialog tracks were identical) and background sounds were minimal at that moment. This audio calibration segment was adjusted to 61 dBA, measured at the viewer's head (ambient noise level of 33 dBA).

4.2. Sweetspot effect

It should be noted that all participants in this study were located at the sweet spot of the rendering system, and they could thus enjoy the best sound reproduction. The impact of an off-axis seating would certainly be more pronounced for the `HYB` soundtrack than it would be for the `WFS` soundtrack as the process of stereo panning relies on the proper positioning of the listener in the sweetspot. Indeed, taking into account the geometry of the reproduction system, the sweet spot of the stereo reproduction has a width of merely 10 cm according to [12]. When outside the sweetspot, sources tend to be attracted to the closer speaker position. On the other hand, the ability of WFS to reproduce a sound position independently from the listener position [13], combined with the ventriloquism effect [14], would result in a larger sweet spot because the sound location is preserved when the listener is off-axis but can still be perceived as coming from the visual object. The congruence in that case is limited by the difference in audio and video perspectives that can be detected by the spectator [15].

4.3. Objective analysis

An objective analysis of the rendered audio was performed. A binaural recording of each condition was made with an artifical head placed at the sweet spot, equivalent to the spectator position during the subsequent experiment. The evolution of the relative sound level at the listener position for the three conditions was measured using a 1 s sliding window and averaged over both ears.

Outside of the region used to calibrate the three conditions, the `ST` soundtrack has a higher level at several moments. This is due to the difference in audio content, as the original track contained a richer audio mix. Some differences are observed between the `WFS` and `HYB` conditions. The different spatialization processes lead to slight differences in sound level that cannot be compensated exactly using only a main volume equalization.

The perceived distribution of the sound sources is of interest. The interaural level differences (ILDs) and the interaural time differences (ITDs) are thus computed from the binaural signals. Binaural signals are subdivided into 1 s segments and analyzed in third-octave bands to obtain ILD and ITD values, using the `Binaural Cue Selection` toolbox [16]. These values are then averaged across pertinent frequency bands (<1.5 kHz for ITD, >1.5 kHz for ILD [17]). The threshold value of 1.5 kHz also corresponds to the SMART-I^2 WFS aliasing frequency.

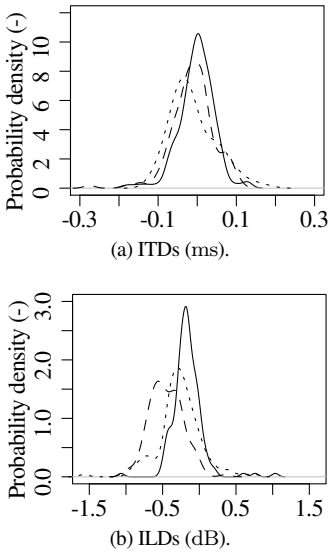


Figure 3: Estimates of the probability density functions of the mean interaural time differences (ITDs) and interaural level differences (ILDs) obtained for each soundtrack. – `ST`, - - `HYB`, ··· `WFS`.

Table 1 presents the mean and standard deviations of the obtained values. In both cases, the mean decreases from `ST` to `WFS` to `HYB`. All means are statistically different from each other, except when comparing the `HYB` and `WFS` ITD means (one-sided Wilcoxon rank sum test, at the 0.05 level). One would also expect that the cues are more spread for `WFS` than for `HYB`. This is the case since the standard deviation increases from `ST` to `HYB` to `WFS` for both ITDs and ILDs.

Histograms of mean ILDs and ITDs are shown in Fig. 3. In both cases, the peak of the probability density function (pdf) is higher for `ST` than it is for `HYB` and `WFS`. This confirms that the `HYB` and `WFS` localization cues are more distributed or spread out than those for the `ST` condition.

5. METHOD

Thirty-three (33) subjects took part in the experiment (26 men, 7 women, age 16 to 58 years, $M = 30.66$, $SD = 10.77$). They answered to a call for participants describing a "3D cinema experiment".

	M_{ITD}(ms)	SD_{ITD} (ms)	γ_{ITD} (-)
ST	-0.0012	0.0445	-0.52
HYB	-0.0112	0.0543	-0.84
WFS	-0.0106	0.0596	0.71
	M_{ILD} (dB)	SD_{ILD} (dB)	γ_{ILD} (-)
ST	-0.16	0.21	1.50
HYB	-0.45	0.22	0.26
WFS	-0.27	0.29	-0.19

Table 1: Means, standard deviations, and skewness of the computed ILDs and ITDs as a function of SOUND CONDITION.

Factor	# questions
Spatial presence	7
Social presence – actor w/i medium	7
Social presence – passive interpersonal	4
Social presence – active interpersonal	3
Engagement (mental immersion)	6
Social richness	7
Social realism	3
Perceptual realism	5

Table 2: The eight components in the Temple Presence Inventory, and the associated number of questions. From [2].

Each was compensated with a soft drink and a cookie while filling the post-session questionnaire.

To determine whether or not the sound modality impacts on the reported sense of presence, a between-subjects experiment was designed. The three different soundtrack conditions, ST, HYB, and WFS (Sect. 4) were used as an independent variable. Each participant was assigned randomly to one particular condition, with 11 participants in each group.

In order to assess the sense of presence as a dependent variable, two methods were used. A post-session questionnaire was developed, providing a subjective assessment. In addition, an oxymeter was used to continuously measure the heart rate of the participants. The goal was to compare this objective measure with the presence score obtained with the questionnaire. The heart rate was measured at 60 Hz using a finger mounted pulse oxymeter (CMS50E, Contec Medical Systems Co.).

It is hypothesized that the spatial rendering quality of sound will impact on the reported sense of presence, as measured by the questionnaire. It is also hypothesized that measures extracted from the heart rate signal will reflect a change from baseline due to the movie presentation and that this change in value is linked to the spatial rendering quality of sound.

5.1. Procedure

Each participant was seated in a comfortable chair (see Fig. 1) in front of the SMART-I^2 and was provided with written instructions regarding the experiment. The oxymeter was placed at the tip of the middle-finger of his/her left hand. The participant was left alone in the experimental room. The room was then completely darkened for a period of 30 s after which the movie was started from a remote control room. This allowed the participant to accommodate him/herself to the darkened environment, and to approach a "cinema" experience. At the end of the movie, the participant was directly taken to the lobby to complete a questionnaire.

5.2. Post-session questionnaires

A presence questionnaire was created using three groups of questions gathered from different sources previously reported. The first group came from the Temple Presence Inventory (TPI) [2], a 42-item cross-media presence questionnaire. The TPI is subdivided into eight groups of questions that measure different aspects of presence. These subgroups, or components, are given in Tab. 2 with the associated number of questions.

The sensitivity of the TPI to both the media form and the media content has been previously confirmed [2]. The second group of questions was taken from the short version of the Swedish Viewer-User Presence (SVUP-short) questionnaire [18]. Three questions regarding the sound rendering were selected. Finally, the last group of questions, which measured negative effects, were from Bouvier's PhD thesis [19].

The resulting questionnaire was translated into French. Each question was presented using a 7-point radio button scale, with two opposite anchors at the extreme values, resulting in a score between 1 and 7. Composite scores were calculated as the mean results for all items in each group.

The principal score of interest is the global score obtained with the TPI, termed TEMPLE. Of all the components in the TPI, the scores *Spatial presence* (SPATIAL) and *Presence as perceptual realism* (PERCEPTUAL_REALISM) are expected to be significantly varying with the media form [2]. The SWEDISH score, from the SVUP-short, gives additional information on the perception of each sound condition. The NEGATIVE score, from Bouvier's PhD thesis, allows one to discard participants who experienced discomfort.

5.3. Heart Rate Variability

Heart Rate Variability (HRV) describes the changes in heart rate over time. Several studies have used HRV as a physiological measure in experiments involving virtual reality [20, 21]. Standards exist [22] describing the different measures that can be extracted from an electrocardiographic (ECG) record. Although HRV is calculated from time intervals between two heart contractions (RR intervals) in an ECG signal, it has been shown that it is possible to obtain the same results from peak-to-peak intervals given by a finger-tip photoplethysmograph (PPG) [23]. Since the signal is captured at only one point on the body, the PPG is less intrusive than the ECG. Analysis of the resulting HRV data was performed in both the time domain and the frequency domain.

The majority of time domain HRV measures require recordings longer than 5 min, which are not possible due to the duration of the film excerpt used. Only the following measures were calculated:

- MeanRR - mean RR interval [ms]
- MinRR - shortest RR interval [ms]
- MaxRR - longest RR interval [ms]
- ΔRR - difference between MaxRR and MinRR [ms]

Frequency domain measures obtained through power spectral density estimation of the RR time series are of particular interest, since their evolution has been correlated with positive or negative emotions when presenting movie clips [24].

In the case of short-term recordings (from 2 to 5 min), three main spectral components are distinguished: the very low frequency (VLF) component between 0.003 Hz and 0.04 Hz, the low frequency (LF) component between 0.04 Hz and 0.15 Hz, and the high frequency (HF) component, between 0.15 Hz and 0.4 Hz. Instead of the absolute values of VLF, LF, and HF power components in ms^2, the values are expressed as LFnorm and HFnorm in normalized units (n.u.), which

represent the relative value of each component in proportion to the total power minus the VLF component.

The parasympathetic activity, which governs the HF power [25], aims at counterbalancing the sympathetic activity, which is related to the preparation of the body for stressful situations, by restoring the body to a resting state. It is believed that LF power reflects a complex mixture of sympathetic and parasympathetic modulation of heart rate [25]. Emotions such as anger, anxiety, and fear, which correspond to the emotions elicited by our movie clip, would be associated to a decreased HF power [26].

6. RESULTS FROM POST-SESSION QUESTIONNAIRES

6.1. Treatment of missing values

There were 10 answers (out of 2785) left blank in the questionnaire results. To avoid discarding the corresponding participants, multiple imputations of the incomplete dataset were used to treat these missing values. This was done using the R [27] package `Amelia II` [28]. Multiple imputation builds m (here five) complete datasets in which each previously missing value is replaced by a new imputed value estimated using the rest of the data. Each imputed value is predicted according to a slightly different model and reflects sampling variability.

In the subsequent analysis, point and variance estimates were estimated according to the method described in [28]. F-statistics and their associated p-value were estimated according to the method given in [29], resulting in analyses of variance (ANOVAs) with degrees of freedom which are no longer integers.

6.2. Negative effects

It is necessary to verify that no participant suffered physically from the experiment. The initial analysis of the results considers the `NEGATIVE` group of questions, measuring negative effects induced by the system, such as nausea, eye strain, or headache.

A bivariate analysis [30] of the `NEGATIVE` score versus the `TEMPLE` score, indicated that one participant was an outlier, reporting feeling much worse than the other participants. This participant was therefore discarded from the study. All others obtained a `NEGATIVE` score less than 2.17 (minimum possible value = 1), which can be considered as having experienced little or no negative effects during the experiment.

6.3. Impact of sound rendering condition on presence

The mean scores in each presence category of interest, obtained for each `SOUND CONDITION`, are given in Tab. 3a. Following an ANOVA analysis, all scores failed to achieve the 0.05 significance level. Hence, no significant effect was observed for sound condition over all subjects.

6.4. A model for the perceived presence

Inspection of the probability density function of `SPATIAL`, `PERCEPTUAL_REALISM`, and `TEMPLE` scores showed them to be non-normal distributions, suggesting a bimodal distribution of two groups centered on different means. This type of distribution can be modeled as a special form of a Gaussian mixture model (GMM). The package `Mclust` [31] allows one to find coefficients of a Gaussian mixture from the data by selecting the optimal model according to the Bayesian information criterion (BIC) applied to an expectation-maximization (EM) algorithm initialized by hierarchical clustering for parameterized Gaussian mixture models.

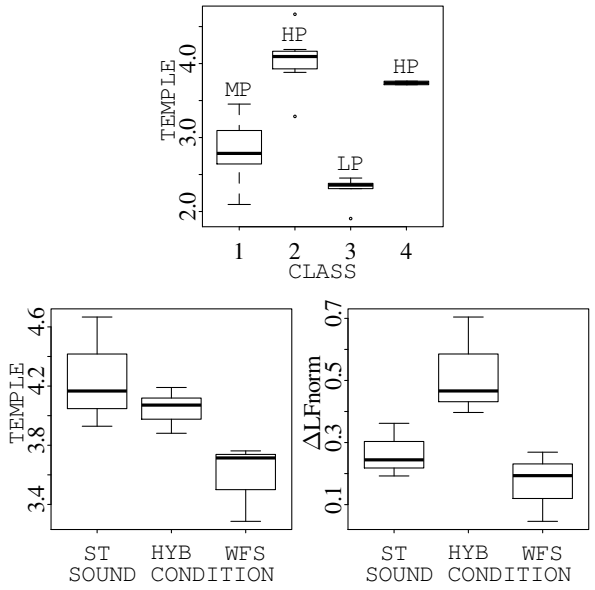


Figure 4: Boxplots of the `TEMPLE` score vs. the GMM classification `CLASS` (top), `TEMPLE` score vs. `SOUND CONDITION` for participants in group `HP` (bottom left), and the ΔLFnorm value vs. `SOUND CONDITION` for participants in group `HP` (bottom right).

The algorithm was run on the data defining the `TEMPLE` score and the resulting optimal model contains four Gaussian components. The probability that a given participant is not correctly classified using this model ranged from 0 to 5.8×10^{-3} ($M = 1.2 \times 10^{-3}$). This demonstrates the good quality of the classification. The four groups, referred to by the factor `CLASS`, are given in descending order of the number of participants they contain: 15, 10, 5, and 2. The mean presence scores for each `CLASS` category are given in Tab. 3b.

Figure 4 (top) shows an analysis of the `TEMPLE` score depending on the classification `CLASS`. Groups 1 and 3 tend to have a lower presence score than the groups 2 and 4. An analysis of variance was carried out on the `TEMPLE` score with the fixed factor `CLASS` (four levels). The factor showed a significant effect ($F_{2.72,27.88} = 39.88$, $p < 10^{-5}$). Subsequent post hoc comparisons (Tukey's HSD test), with an α level of 0.05 showed that groups 2 and 4 do not differ significantly (they form a homogeneous subset), while groups 1 and 3 are significantly different and both differ from the aforementioned set of groups 2 and 4. In the following sections, group 3 will be referred to as `LP` (low presence, 5 subjects), group 1 as `MP` (medium presence, 15 subjects), and the combination of groups 2 and 4 as `HP` (high presence, 12 subjects).

6.5. Further analysis in each group

An analysis of variance was carried out on the `TEMPLE` score with the fixed factor `SOUND CONDITION` (three levels) for each presence group defined in the previous section. The factor showed a significant effect on group `HP` ($F_{1.95,8.99} = 6.85$, $p = 0.016$). However, `SOUND CONDITION` was significant neither for group `MP` ($F_{2.00,11.90} = 0.11$, $p = 0.896$) nor for group `LP` ($F_{1.00,3.00} = 0.69$, $p = 0.468$).

Subsequent post hoc comparisons (Tukey's HSD test, $\alpha = 0.05$) on the group `HP` showed that conditions `ST` (4 participants) and `HYB`

	ST	HYB	WFS	F	p-value
SPATIAL	2.90 (1.2)	2.83 (1.12)	2.47 (0.73)	0.50	0.900
PERCEPTUAL_ REALISM	2.71 (1.1)	2.71 (0.87)	2.66 (0.74)	0.01	0.991
TEMPLE	3.34 (0.79)	3.23 (0.87)	2.94 (0.54)	0.79	0.487
SWEDISH	5.15 (0.78)	4.97 (1.34)	4.57 (0.83)	0.89	0.773

(a) SOUND CONDITION

	1	2	3	4	F	p-value
SPATIAL	2.25 (0.57)	3.81 (0.78)	1.71 (0.29)	3.64 (0.51)	19.49	$< 10^{-5}$
PERCEPTUAL_ REALISM	2.08 (0.46)	3.66 (0.49)	2.40 (0.91)	3.20 (0.57)	17.07	$< 10^{-5}$
TEMPLE	2.82 (0.39)	4.05 (0.34)	2.28 (0.22)	3.74 (0.03)	39.88	$< 10^{-5}$
SWEDISH	4.78 (0.97)	5.73 (0.73)	4.00 (0.53)	4.00 (0.00)	6.24	0.107

(b) CLASS

Table 3: Presence questionnaire scores for each category by (a) SOUND CONDITION and (b) CLASS (means and standard deviations).

(5 participants) do not differ significantly (they form a homogeneous subset), while condition WFS (3 participants) differ significantly from the conditions ST and HYB.

Figure 4 (bottom left) shows an analysis of the TEMPLE score for each sound condition in group HP. Presence scores are (statistically) lower in the WFS group than in the two other groups. A similar analysis was performed on the SPATIAL, PERCEPTUAL_REALISM, and SWEDISH scores. These failed to achieve the 0.05 significance level. This result, combined with the result on the TEMPLE score, indicates that the impact of sound reproduction is spread accross different components of presence rather than confined to the components *Spatial presence* and *Perceptual realism*.

In summary, the sound condition does not affect the reported presence score directly for all subjects. Rather, participants can be classified according to their presence score independently of the sound condition. In the group that reported the highest sense of presence, for which sound rendering condition was influential, the spatially coherent soundtrack (WFS) is significantly different from the two other stereo soundtracks. The WFS soundtrack leads to a decreased reported sense of presence.

6.6. Discussion

SOUND CONDITION as an independant variable fails at predicting the obtained presence score for all participants. Rather, the participants are classified according to their presence score in three groups. The first has a low presence score (LP), the second has a somewhat higher presence score but also a higher variability (MP), and the third has a high presence score (HP).

SOUND CONDITION has a statistically significant impact for the group HP. In this group, the HYB soundtrack is not statistically different from the original ST version, which means that the slight difference in content between the two soundtracks did not impact on the reported sense of presence.

When comparing the results for the HYB and the WFS soundtracks, one can see that there is a statistical difference in reported sense of presence which is to the advantage of HYB. In this condition, sound objects were limited to the space between the virtual speakers, and since the participants were at the sweet spot, objects in-between were fairly well localized in azimuth. Therefore, one could hypothesize that presence is lessened when the auditory objects extend beyond the screen boundaries. Indeed, the virtual loudspeakers in the HYB condition were located near the screen borders, and Fig. 3 shows the spread of the mean ITDs increasing with SOUND CONDITION, from ST to WFS. Further studies with different source material would be required to substantiate this hypothesis.

HRV	Baseline	Experiment	p-value
MeanRR [ms]	835.8	849.7	0.026
MinRR [ms]	648.9	627.4	0.044
MaxRR [ms]	1024.2	1135.5	0.007
ΔRR [ms]	375.3	508.1	0.006
LFnorm [n.u.]	42	55	0.016
HFnorm [n.u.]	58	45	0.016
LF/HF [/]	1.08	1.75	0.034

Table 4: HRV time and frequency domain parameters

7. RESULTS FROM HEART RATE VARIABILITY

The analysis presented in the previous section is repeated here on the recorded heart rate, using the same statistical software. Due to a technical glitch, however, the heart rate could not be recorded for one of the participants, who is thus not included.

7.1. Overall comparison of baseline and experimental phases

Table 4 shows the HRV parameters averaged over all subjects for the two phases: *baseline*, when the participant is in the dark, and *experiment*, when the participant watches the movie. Since the data does not meet the normality assumption, a non-parametric test, the Wilcoxon signed rank test, was applied between the parameters of the baseline and the experiment. The values of the four parameters are statistically different, at the 0.05 level, between the two phases.

Table 4 shows the changes of HRV parameters in the frequency domain averaged over all subjects for the two same phases. The Wilcoxon signed rank test was applied between the parameters of the baseline and the experiment. The last column gives the corresponding p-values. All the HRV frequency parameters are statistically different at the 0.05 level.

In agreement with the litterature [32], HRV allows one to discriminate between rest and "work" (the movie presentation). The decreasing HF component is similar to that observed in [24] where different positive and negative emotions are expressed through different movie clips.

7.2. Heart Rate Variability

To investigate the effect of SOUND CONDITION on HRV, an analysis of variance was carried out on the difference between LFnorm during experiment and baseline (ΔLFnorm) with the fixed factor SOUND CONDITION (three levels) for each presence group defined in Section 6.4. The factor showed no significant effect on any group at the 0.05 level. However, the factor showed a significant effect on

	Sample estimate	t_{29} (t_{28})	p-value	Lower bound	Upper bound
SPATIAL	0.41 (0.49)	2.41 (2.95)	0.022 (0.006)	0.06 (0.15)	0.67 (0.72)
PERCEPTUAL_ REALISM	0.42 (0.44)	2.47 (2.56)	0.020 (0.016)	0.07 (0.09)	0.67 (0.69)
TEMPLE	0.45 (0.52)	2.73 (3.23)	0.011 (0.003)	0.12 (0.20)	0.70 (0.74)
SWEDISH	0.52 (0.56)	3.24 (3.54)	0.003 (0.001)	0.20 (0.24)	0.74 (0.76)

Table 5: Pearson's product-moment correlation between ΔLFnorm and the presence scores (in the first imputed dataset). In parentheses, the values obtained when participant 2 is discarded.

the group HP ($F_{2.00,7.00} = 7.68$, $p = 0.017$) if participant 2 was removed from the analysis. According to a bivariate analysis [30], participant 2 would not be classified as an outlier, though he is near the limit. Still, this subject was the only one to exhibit a negative ΔLFnorm (decrease relative to the baseline). As such, further results have been calculated both with and without subject 2 included.

Subsequent post hoc comparisons (Tukey's HSD test, $\alpha = 0.05$) on the group HP showed that conditions ST (4 participants) and WFS (3 participants) do not differ significantly (they form a homogeneous subset), while condition HYB (3 participants) differs significantly from this set of conditions.

Figure 4 shows analysis of ΔLFnorm values for each sound condition in group HP. ΔLFnorm values are (statistically) higher in the HYB group than in the two other groups. Similar results can be obtained with ΔHFnorm, since it is linearly dependent on ΔLFnorm. The results obtained with ΔLF/HF fail to reach the 0.05 significance level as well as the results obtained with the time-domain parameters.

In summary, the ideal stereo version (HYB) is significantly different from the two other soundtracks in the group of subjects that reported the highest sense of presence. The HYB soundtrack leads to an increased low frequency component of the HRV.

7.3. Relationship between HRV and questionnaire scores

In order to evaluate the correlation between the questionnaire scores and the evolution of the frequency-domain HRV parameters, Pearson's product-moment correlation was computed. The results, including the 95% confidence interval, are presented in Tab. 5. Naturally, the opposite values are found for ΔHFnorm.

The correlation is significantly different from 0 (at the 0.05 level) for every presence score of interest. The highest value is obtained with the SWEDISH score, which pertains only the sound rendering. When participant 2 is discarded from the analysis, the values are improved. This is indicated in parentheses in Tab. 5.

7.4. Discussion

The presentation of the movie to the participants had an impact on several Heart Rate Variability (HRV) statistics in both time and frequency domains. For all participants, a relation is found between the reported presence score TEMPLE and the evolutions of both LFnorm and HFnorm between the baseline and the experiment.

SOUND CONDITION as an independent variable fails at predicting the obtained evolutions of HRV parameters for all participants. The analysis according to each presence group shows that SOUND CONDITION has a statistically significant impact on ΔLFnorm and ΔHFnorm for the group HP (with participant 2 discarded). In that case, the HYB soundtrack is statistically different from both the original ST version and the WFS version.

When comparing the HYB and the WFS soundtracks, one can see that there is a statistical difference in the evolutions of LFnorm and HFnorm, which is higher in the HYB case. Participants in the HYB condition therefore experienced a higher increase in LFnorm than the others. Since ΔLFnorm correlates positively with TEMPLE for all participants, this supports our previous findings that the participants experienced a stronger sense of presence with the HYB soundtrack than with the WFS soundtrack.

8. RESULTS FROM THE PARTICIPANTS' FEEDBACK

Among the comments the participants made on the experiment, a few recurring ones can be outlined. Nine participants indicated that they were disappointed by the (visual) 3D. Maybe they expected to see more depth in the movie than they actually saw. As can be seen in Fig. 2, the range of depth of the sources is rather narrow (roughly 0.5 m to 5 m). The length of the experience was also a problem for seven participants who reported it being too short. They needed more time to forget they were in an experiment. Five participants found the end of the movie excerpt too abrupt, they would have appreciated to know more about the story. Regarding the setup, four participants were distracted by the visibility of the corner of the panels in the SMART-I^2 and three complained about the glasses (two of which wore prescription glasses).

It is therefore possible that the results found in this study could vary, or be improved, if a longer film was shown, and if the projection was made on a traditional flat format screen. These comments will be taken into consideration in future studies.

The comments made by the participants underline the limitations of this experiment. Most were related to the content, rather than the setup. Some participants found that the movie did not present much depth, and that the movie was too short to allow some of them to forget they were taking part in an experiment. Several participants were disappointed with the end of the story, or even did not like the movie at all.

9. CONCLUSIONS

Different sound spatialization techniques were combined with an s-3D movie. The impact of these techniques on the sense of presence was investigated using a post-session questionnaire and heart rate monitoring.

The sound condition did not affect the reported presence score directly for all subjects. Rather, participants could be classified according to their presence score independently of the sound condition. In the group that reported the highest sense of presence, for which sound rendering condition was influential, the spatially coherent soundtrack (WFS) was significantly different from the two other stereo soundtracks. The WFS soundtrack led to a decreased reported sense of presence. Analysis of the participants' Heart Rate Variability (HRV) revealed that, in the group that reported the highest sense of presence, the ideal stereo version (HYB) was significantly different from the two other soundtracks. The HYB soundtrack led to an increased low frequency component of the HRV.

The HRV low frequency component was also shown to be positively correlated to the overall presence score for all participants. Both the subjective (questionnaire) and objective (HRV) measures showed that the HYB soundtrack led to a higher sense of presence than the WFS one for participants that reported the highest sense of presence.

The results found here constitute a basis for future research. The impact of an off-axis seating position needs further investigation, since the

s-3D image is egocentric. Apart from the reverberation, all the sound in this experiment came from the front. Therefore, there is also a need to investigate the effect with a full 360° sound reproduction. Finally, one could investigate other types of 3D sound rendering, such as Ambisonics, binaural, or possible hybrid combinations of multiple systems.

10. REFERENCES

[1] I. Allen, "Matching the sound to the picture," in *Audio Eng. Soc. 9th Int. Conf.: Television Sound Today and Tomorrow*, 1991.

[2] M. Lombard, T. Ditton, and L. Weinstein, "Measuring (tele)presence: The Temple Presence Inventory," in *12th Int. Workshop on Presence*, Los Angeles, CA, 2009.

[3] W. Ijsselsteijn, H. de Ridder, J. Freeman, S. E. Avons, and D. Bouwhuis, "Effects of stereoscopic presentation, image motion, and screen size on subjective and objective corroborative measures of presence." *Presence-Teleop. Virt.*, vol. 10, no. 3, pp. 298–311, 2001.

[4] M. Rébillat, E. Corteel, and B. F. G. Katz, "SMART-I^2: Spatial Multi-User Audio-Visual Real Time Interactive Interface," in *Audio Eng. Soc. Conv. 125*, 2008.

[5] M. M. Boone, "Multi-Actuator Panels (MAPs) as Loudspeaker Arrays for Wave Field Synthesis," *J. Audio Eng. Soc.*, vol. 52, no. 7/8, pp. 712–723, 2004.

[6] A. J. Berkhout, "A Holographic Approach to Acoustic Control," *J. Audio Eng. Soc.*, vol. 36, no. 12, pp. 977–995, 1988.

[7] É. Corteel, "On the use of irregularly spaced loudspeaker arrays for Wave Field Synthesis, potential impact on spatial aliasing frequency," in *Proc. 9th Int. Conf. on Digital Audio Effects (DAFx'06)*, Montréal, Canada, 2006.

[8] J.-M. Jot, "Real-time spatial processing of sounds for music, multimedia and interactive human-computer interfaces," *Multimedia Systems*, vol. 7, no. 1, pp. 55–69, 1999.

[9] "Spat reference manual," http://forumnet.ircam.fr/692.html.

[10] S. Eilemann, M. Makhinya, and R. Pajarola, "Equalizer: A scalable parallel rendering framework," *IEEE Trans. Vis. Comput. Graphics*, vol. 15, no. 3, pp. 436–452, 2009.

[11] M. Évrard, C. R. André, J. G. Verly, J.-J. Embrechts, and B. F. G. Katz, "Object-based sound re-mix for spatially coherent audio rendering of an existing stereoscopic-3D animation movie," in *Audio Eng. Soc. Conv. 131*, New York, NY, 2011.

[12] G. Theile, "On the performance of two-channel and multi-channel stereophony," in *Audio Eng. Soc. Conv. 88*, 1990.

[13] G. Theile, H. Wittek, and M. Reisinger, "Potential wavefield synthesis applications in the multichannel stereophonic world," in *Audio Eng. Soc. 24th Int. Conf.: Multichannel Audio, The New Reality*, 2003.

[14] W. R. Thurlow and C. E. Jack, "Certain determinants of the "ventriloquism effect"," *Percept. Motor Skill*, vol. 36, pp. 1171–1184, 1973.

[15] W. P. J. de Bruijn and M. M. Boone, "Subjective experiments on the effects of combining spatialized audio and 2D video projection in audio-visual systems," in *Audio Eng. Soc. Conv. 112*, 2002.

[16] C. Faller and J. Merimaa, "Source localization in complex listening situations: Selection of binaural cues based on interaural coherence," *J. Acoust. Soc. Am.*, vol. 116, no. 5, pp. 3075–3089, 2004.

[17] J. Blauert, *Spatial hearing: the psychophysics of human sound localization*. MIT Press, 1997.

[18] P. Larsson, D. Västfjäll, P. Olsson, and M. Kleiner, "When what you hear is what you see: Presence and auditory-visual integration in virtual environments," in *Proc. 10th Annu. Int. Workshop Presence*, Barcelona, Spain, 2007, pp. 11–18.

[19] P. Bouvier, "La présence en réalité virtuelle, une approche centrée utilisateur," Ph.D. dissertation, Université Paris-Est, Paris, France, 2009.

[20] M. Slater, C. Guger, G. Edlinger, R. Leeb, G. Pfurtscheller, A. Antley, M. Garau, A. Brogni, and D. Friedman, "Analysis of physiological responses to a social situation in an immersive virtual environment," *Presence-Teleop. Virt.*, vol. 15, no. 5, pp. 553–569, 2006.

[21] S. Huang, P. Tsai, W. Sung, C. Lin, and T. Chuang, "The comparisons of heart rate variability and perceived exertion during simulated cycling with various viewing devices," *Presence-Teleop. Virt.*, vol. 17, no. 6, pp. 575–583, 2008.

[22] Task Force of The European Soc. of Cardiology and The North Am. Soc. of Pacing and Electrophysiology, "Heart Rate Variability: Standards of measurement, physiological interpretation, and clinical use," *Circulation*, vol. 93, no. 5, pp. 1043–1065, 1996.

[23] N. Selvaraj, A. Jaryal, J. Santhosh, K. K. Deepak, and S. Anand, "Assessment of heart rate variability derived from finger-tip photoplethysmography as compared to electrocardiography," *J. Med. Eng. Technol.*, vol. 32, no. 6, pp. 479–484, 2008.

[24] E. Vianna and D. Tranel, "Gastric myoelectrical activity as an index of emotional arousal," *Int. J. Psychophysiol.*, vol. 61, no. 1, pp. 70–76, 2006.

[25] P. Stein and R. Kleiger, "Insights from the study of heart rate variability," *Annu. Rev. Med.*, vol. 50, no. 1, pp. 249–261, 1999.

[26] S. D. Kreibig, "Autonomic nervous system activity in emotion: A review," *Biol. Psychol.*, vol. 84, no. 3, pp. 394–421, 2010.

[27] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2011.

[28] G. King, J. Honaker, A. Joseph, and K. Scheve, "Analyzing incomplete political science data: An alternative algorithm for multiple imputation," *Am. Polit. Sci. Rev.*, vol. 95, pp. 49–69, 2001.

[29] T. Raghunathan and Q. Dong, "Analysis of variance from multiply imputed data sets," Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI, Tech. Rep.

[30] K. M. Goldberg and B. Iglewicz, "Bivariate extensions of the boxplot," *Technometrics*, vol. 34, no. 3, pp. 307–320, 1992.

[31] C. Fraley and A. E. Raftery, "Enhanced model-based clustering, density estimation, and discriminant analysis software: MCLUST," *J. Classif.*, vol. 20, pp. 263–286, 2003.

[32] P. Nickel and F. Nachreiner, "Sensitivity and diagnosticity of the 0.1-Hz component of heart rate variability as an indicator of mental workload," *Hum. Factors*, vol. 45, no. 4, pp. 575 –590, 2003.

EXPLORING 3D AUDIO FOR BRAIN SONIFICATION

Timothy Schmele

Barcelona Media
Av. Diagonal 177
Barcelona, Spain
tim.schmele@barcelonamedia.org

Imanol Gomez

FutureLab
Ars Electronica
Linz, Austria
Imanol.Gomez@aec.at

ABSTRACT

Brain activity data, measured by functional Magnetic Resonance Imaging (fMRI), produces extremely high dimensional, sparse and noisy signals which are difficult to visualize, monitor and analyze. The use of spatial music can be particularly appropriate to represent its contained patterns. The literature describes several research done on sonifying neuroimaging data as well as different techniques to use spatialization as a musical language. In this paper, we discuss an artistic approach to fMRI sonification exploiting new compositional paradigms in spatial music. Therefore, we consider the brain activity as audio base material of a the spatial musical composition. Our approach attempts to explore the aesthetic potential of brain sonification not by transforming the data beyond the recognizable, but presenting the data as direct as possible.

1. INTRODUCTION

Functional Magnetic Resonance Imaging (fMRI) provides the user with information on the location of functional activations in the different regions of the brain with high spatial resolution. The resulting data is highly dimensional, sparse and noisy, and is difficult to monitor and detect structures or patterns. This fact has motivated the approach to improve the exploratory data analysis. The main goal is to use sound to render the original data in a suitably transformed way, so that we can invoke our natural pattern recognition capabilities to search for regularities and structures.

In particular, these capabilities and mechanisms are triggered involuntarily during the act of listening to what we perceive as music. When listening to music, the brain constantly estimates the continuation of a musical gesture. We find pleasure in the encounter of a musical pattern and so this search for connections and an apparent message in music comes natural to us. At the same time, interest needs to be maintained by providing surprises and unforeseen developments that make us reconsider our previous estimations keeps the music engaging. The main job of a composer is to skillfully play with this expectation and keep up the interest by violating the predictions made and breaking the patterns.

Sonification in music makes use of patterns contained in the data to be sonified. A composer of algorithmic music consciously takes the decision to step back from his foremost compositional responsibilities and lets the algorithm and the data take control of the musical creation to a large part. Algorithmic composition requires human intervention on higher, more abstract levels [1]. Decisions such as the proper mapping of parameters, processing and filtering of inaudible data need to be made, while the minor details are left to chance. Listening to this style of music may serve both an aestehtic and scientific purpose. For their database of *Sonification in Music*, Schoon and Dombois [2] define three criteria for inclusion of a work: the transformation from inaudible to audible frequency, the acquisition of knowledge through the act of listening, as well as the development of listening techniques that are subject to scientific validation.

In this paper, we discuss an artistic approach to fMRI sonification that exploits new compositional paradigms in spatial music, attempting to establish the physical space around the listener as a musical language of its own. That is, beyond the ability to utilize frequency, rhythm and timbre among other musical parameters, the process of spatializing music is not just a tool for further clarification of the sonic material, but part of the compositional process and is considered musical gesture in itself. In a sense, a sonorous gesture in physical space is comparable to a melody and closely linked to timbre and rhythm.

Even though the human hearing system is known to be able to decode and interpret complex auditory scenes [3], the more structured the representation of the sonified data, the better the accessibility and intelligibility of the chosen process. Hence, presenting both distinct data and interpretations of the data in respective, designated musical dimensions aids in bringing clarity to the audible scene. Adding the ability to spatialize music in full, continuously and freely moveable three dimensional space opens new possibilities to data sonification and changes the way sounds are interpreted in relation to their perceived spatial location.

2. BACKGROUND

2.1. Sonification for data exploration

With abundance of high-dimensional data, auditory data exploration has become an important tool to comprehend such data and to uncover its structures and patterns [4, 5]. Thus, sonification has expanded beyond the classic process monitoring applications and many researchers among different fields are currently researching in this area.

Vogt et al. [6] used sonification to understand lattice quantum chromodynamics (QCD) as a representation of a 4 dimensional space; Grond et al. [7] implemented a combined auditory and visual interface to help browsing ribonucleic acid (RNA) structures; Winters et al. [8] simulated through sound the phase transition that occurred shortly after the Big Bang; Bearman [9] used sound to represent uncertainty in future climate predictions; Alexander R. et. al [10] was able reveal new insights into data parameters for differentiating solar wind types, by audifying and listening to 13 years of heliospheric measurements.

Sonification is particularly appropriate to improve the understanding of neuroimaging data, which is naturally multidimensional. There have been several studies that have focused on analysing the data obtained from Electroencephalography (EEG) measurements. One of the first attempts to auditory EEG exploration was reported in 1934 by E. Adrian and B. Matthews [11]. For their research they measured the brain activity from a human subject by electrodes applied to the head, and the channels were viewed optically on bromide paper using the Matthews oscillograph, while being directly transduced into sound. More recently, T. Hermann et al. have presented different strategies of sonification for human EEG [12, 13, 14, 15] and Gomez et al. [16] studied different approaches to fMRI brain data sonification.

Music has also been used to represent human EEG. One example is the work of D. Wu et al., representing mental states by using music [17]. The EEG features were extracted by wavelet analysis and they would control musical parameters such as pitch, tempo, rhythm, and tonality. To give more musical meaning, some rules were taken into account like harmony or structure. One of the main challenges of this work was to find the precise trade-off between direct sonification of the features and music composition.

One of the most relevant musical outcomes was the concert of sonification at the Sydney Opera House, for the ICAD 2004 [18]. Ten pieces of music were composed from an EEG data set of a person listening to a piece of music. Whilst performed the audience stood immersed during the concert in a 16.2 dome of speakers arranged to mimic the positions of EEG electrodes on the scalp. Although most participants made use of the speaker configuration, the musical impact of placing specific sound material in each respective location is rarely discussed. Sonically, section 1 of the piece *The Other Ear* by John A. Dribus shows similarities, in the sense that he creates a fast swirling sensation to represent the brain's activity.

2.2. Spatialization as a musical language

Space is present in most musical vocabulary, as well as projected into many other musical characteristics and parameters. All acoustic instruments have physical dimensions that place certain pitches to unique physical locations. Not just because of this is pitch mostly described with being *high* or *low*; we naturally associate high frequencies as coming from above and vice versa [19]. Moreover, the term 'space' is used in many musical contexts besides meaning actual physical space. Musicologists may refer to tonality as *pitch space*, or to orchestration as *timbral space* [20]. In his writing on space-form and the acousmatic image [21], Smalley presents a new musical taxonomy to compliment qualities specific to electro-acoustic music and bases is completely on the notion of space in music. The spatial development of a sound and its timbral development, the *spectromorphology* as he coins it, become one. Hence, the interpretation of a sound's more traditional audible qualities and the space it occupies are fused together.

Space and the concept of spatialization in electronic music today is a substantiated aspect of the music and is used in a unique and radically different way compared to an previous acoustic effort [22, 21, 23]. As Normandeau points out, the development of the loudspeaker had a fundamental impact on the way composers see space [23]. Being able to play any sound or timbre, especially sharing the exact same signal as another loudspeaker, makes this electronic device a unique instrument. If two loudspeakers play an identical sound the brain will fuse these two signals together, making it appear for this single sound be coming from the space between the speakers. An imbalance of amplitude between the speakers moves the sound from one speaker to another and makes the space in which the sound can travel *continuously*. Hence, the realization that a massless, virtual sound source may travel at virtually any speed to any place had a significant impact of musical thinking in the 20th century.

While the above technique, also known as stereo panning, is based on how the brain combines the auditory signals coming from both ears, not all spatialization technologies make use of these psychoacoustic principles. *Wave Field Synthesis* (WFS), in particular, tries to reconstruct the original wavefront of the virtual source from speaker array onwards [24]. Unfortunately, this requires a large amount of speakers and exhibits spatial aliasing above a certain frequency, depending on the size and proximity of the speakers. *Ambisonics* is another sound field reconstruction method but driven by psychoacoustic amplitude panning techniques, similar to stereophony [25]. Compared to *Vector Based Amplitude Panning* (VBAP), it has with a more uniform phantom image but suffers from spatialization blur [26]. In turn, VBAP, being more closely related to stereophony, triangulating the signal between the three nearest speakers [25], demonstrates a higher positioning accuracy.

2.2.1. Cultural developments in spatial music

As of the 20th century, the spatialization of music has received much focus since the dawn of the modernist period, especially with technological advances in sound reproduction techniques and electro-acoustic music on the music's increasing popularity around the 1950's [22, 27]. But the notion of space in musical composition goes back farther than one might suspect at first. Traces can be found starting from the deliberate separation of ensemble parts to articulate antiphonal compositions in biblical times [28], continuing with architecturally motivated compositions, over symbolical spaces and up to virtual soundscapes.

Around the 16th century, antiphonal psalmody heightened with the popularization of the polychoral style, specifically in Venice. The architecture of the venetian Basilica San Marco, with its two spatially separated choir lofts, is said to have inspired composers Adrian Willaert and, most famously, Giovanni Gabrieli to make impressive use of a technique known as *cori battente* or *cori spezzati* for dramatic spatial effects [26, 28, 29]. Although the use of space played an important role in their music, exact spatial arrangements were usually not indicated in the score [22]. It was usually separate the individual groups spatially, meaning that space was merely an implement for a heightened experience as opposed of true compositional concern.

While composers of the classical period showed little interest in spatial effects, there were notable exceptions, however. Wolfgang Amadeu Mozarts Serenada Notturna (1776) for two small orchestras and Notturno (1777)[1] for four Orchestras, demonstrate a tight interweaving of physical space with the music through motivic segmentation and dynamic interplay. He creates echo effects by not simply repeating phrases with each respective orchestra delayed in time, but considers dynamics, masking effects and gradually adds mutes to more instruments in each repetition to denote a gradual darkening at each reflection [29]. Later on, romantic composers would utilize spatial effects for programmatic

[1] Mozart's quadrophonic orchestra piece may sometimes be (strictly speaking, incorrectly) labeled Serenade, such as it is the case in [29] and [26]

purposes, such as the apocalyptic trumpets in Hector Berlioz' *Requiem* (1837) [30], or the use of off stage ensembles, as it is the case in the *finale* in Gustav Mahler's *Symphony No. 2* [26].

Having composed more than half his catalogue of work with deliberate spatial intentions, Henry Brant was one of the first to base his compositional methodology around the musical potential of space. His main concern was the clarification of dense textures through spatial separation [31]. He would mainly approach this problem by spatially separating the instruments into timbral groups to achieve the highest sonic distinction and prevent an effect similar to stereo panning. Furthermore, seating plans were often precisely indicated, which made his compositional techniques possible, such as trajectories and *travel and filling-up*, a gradual engulfment in sound by successively adding instruments to the overall sounding cluster.

But it was not until the introduction of the loudspeaker that the use of space in music was completely revolutionized. With the absence of harmony in atonal music and the replacement of pitch by concrete sounds in the first half of the 20th century, composers were in need of other musical parameters to communicate their compositional intentions. Edgar Varèse thought of sound as a musical object that "[...]flow, change, expand and contract, yet they have a certain tangibility, a concreteness established by clearly defined boundaries." [22]. For the Phillips Pavilion at the 1958 Brussel World Fair, he used an estimated 350 speakers to create sonic trajectories as a central element to his specifically composed *Poème Électronique* [32].

For Karlheinz Stockhausen, the spatial parameter was an inherent part of a sound and was fully integrated into *Total Serialism*. His acclaimed composition *Gesang der Jünglinge* (1956) was originally written for six channels and the serial spatialization of the sound is said to be the most fascinating features [33]. He created both electronic and orchestral pieces with clear spatial intentions in mind, such as *Gruppen* (1955-57) for three orchestras. For the Osaka World Fair in 1970, Stockhausen built the first fully spherical concert hall and created the ability to spatialize sound freely in all dimensions, even below the listener. He divided the space vertically into layers, which were individually treated in the score, with specific interpolative symbols between them [34].

Contemporary trends in spatio-musical composition turn away from a mere trajectory-oriented thinking look more at space itself as a compositional mean. During the performance of *HP-SCHD* (1967-69), John Cage forced the listener to use his directional hearing and decide what to listen to by bombarding him with sounds during a "[...]five hour multi-media extravaganza[...]" [22]. Alvin Lucier famously made space his instrument in *I am sitting in a room* (1969), amplifying the rooms resonant frequencies by successively projecting and re-recording an initial phrase. Kerry Hagan, in turn, engages in textural composition [35], creating new, imginary spaces by engulfing the listener with stochastically placed granules. Putting the listener into the role of the composer, Ryoji Ikeda plays with the perception of space in *db* (2012), as the projectied composition of sine tones through a parabolic speaker is modified through ones own movement in the Hamburger Bahnhof, Berlin, as well as the reflections of other visitors that walk through the sonic beam.

Lastly, Smalley [21], already mentioned above, recognizes the ability of space to change the sounds spectromorphology. Space is not just a parameter the composer can change at will, one needs to be aware of the impact it has on the sound and the changes that happen to the actual music. Smalley coined the term *spatiomorphology*, referring to space as an appreciative experience in itself. He distinguishes spatiomorphology from using space only as means to enhance the spectromorphology. Simply put, this is where he delineates space from being a mere effect as opposed to a parameter suitable for musical expression.

2.2.2. Perception of spatio-musical gestures

Spatial listening is often dealt with the well known binaural cues that describe our ability to make use of our spatially separated ears and shape of our cochlear. But differences in time, level and spectral content are only half the truth. The localization models usually consider an isolated part of the frequency spectrum and would relate to real world situations only if the brain would receive a single anechoic source. Instead, our ears are constantly bombarded with many different sounds from all directions simultaneously.

To separate and localize cohesive, individual entities in this frequency agglomerate coming in through two small openings in our head, Bregman formulated a theory called *Auditory Scene Analysis* (ASA) [36]. Its essence builds on the five founding principles of Gestalt theory around which the theory of grouping and segregation, separating the figure from the ground, are formed [37]: *Similarity*, *Proximity*, *Continuity*, *Common Fate* and *Symmetry & Closure*. Segregation is caused through contrast. Two objects separate one from another not from their relation to each other, but in their relation to their background. For this, Bregman [36] defines a perceptual distance d, that describes a weighted distance between several comparative auditory dimensions, such as frequency or time.

ASA is based on two auditory grouping phenomena: Primitive segregation describes our natural abilities to segregate sounds in the environment from one another, similar to how Gestalt theory describes the urge to see patterns. Spatial cues are a major component in the process of primitive segregation and include both spatial location and spatial continuity among other cues. [37]. *Schemas* come into play where primitive segregation fails, as an additional model of learning, a way of discerning learned patterns from previous events that involved attention and may regroup previously, primitively segregated scenes.

But, beyond ASA, lie higher levels abstractions of our spatial perception. Listening to sounds in space is not fulfilled until we create a mental map of the auditory scene that we may then interpret. Phenomenology, for example, calls for time being the main mediator of this experience and the notion of space as a personal, egocentric perception with movement being the essential bodily experience [22]. This means that spatial perception – spatial awareness is not just individual, but acquired and learnable.

Even though the identification of spatial gestures as a musical act might be alien to some, the musical intentions behind the spatialization of Varèse and Stockhausen, for example, may be understood, if not personally, then culturally, on a larger time scale. Cage and Ikeda, for example, deliberately turn the focus onto the space by reducing other parameters either through overload or reduction. Music that is primarily concerned with space, but fails to address the spatial engagement will be completely misinterpreted. This form of reduced listening can be compared to that proposed in *musique concrète* [38]: aural spatial perception lies within this (usually) subconscious realm of *detectability* [36]. By putting the listener into a reduced state of mind, the composer may push his intentions into the categories of *perceptability* and *desirability* [38] and engages the listener in *attentive listening* [39].

While, at first, spatial music may seem as if it is a pure sensory experience, one just needs to look at visitors that stands in awe of the auditory space of a cathedral, the reverberation and the soundscape of small footsteps in the distance, the mumbling of soft prayers, the occasional camera clicking away. "In many situations, listeners may not be consciously aware of the affect induced by listening to engaging sound or spaces." [39]. Through reduction of other musical parameters the audience has to come to a conclusion that it was not the sounds that moved them – it had to be the space. The composer can steer the the attention to shift the listener from the *detection* of space to the attentive mode of *perception*, but the language of [...] high-impact, emotionally engaged listening [39] can only come from a rich pool of culturally established norms – and a true musical spatial language is still to be established.

3. BRAIN DATA

All the data used for this article was created during the experiments done by Grahn and Rowe in 2009 [40]. In their work they used fMRI images to study the perception of rhythm in musicians and non musicians. In their experiments, several subjects had their brain activity measured, while exposed to volume accented and duration accented rhythmic stimuli.

Every brain image obtained, contained thousands of "voxels" (Volumetric Picture Element), that have been filtered to reduce random noise in the image improves the ability of a statistical technique to detect real activations and reject false ones. Spatially smoothing each of the images improves the signal-to-noise ratio (SNR), as well as temporally smoothing avoids a number of slow "scanner drifts".

fMRI data has a lot of features and fewer examples. Hence, it is desirable to reduce the number of features using feature selection techniques. For our purpose the voxels will be the features to extract. We want to know "how important the voxels of a certain region are, according to the task. The strategy used is voxel discriminability. For each voxel and considered cognitive state, an analysis of variance (ANOVA) is performed comparing the fMRI activity of the voxel in examples belonging to the different stimuli of interest. More concretely, the method chosen is the *one-way analysis of variance*, with a test statistic called *F ratio*. A certain number of voxels can be now selected by choosing the ones with larger f-values. More detail information about the data extraction is described in [16].

Finally, the extracted features are projected onto a hemisphere through a line joining the center of the brain to a point on the surface, and intersecting the top half of a circumscribed sphere (Figure 1).

4. BRAIN AESTHETICS

In order to sonify the extracted features (section 3) into music, we have taken several aesthetic considerations and various levels of abstraction. We want to bring harmony to the formal features, while revealing new insights into reality. The dimensionality of the brain and its activity in terms of voxel energy should be directly perceivable. It is a deliberate choice to turn the brain into a musical instrument by presenting the data as directly as possible. The intention is to explore the aesthetic potential not by transforming the data beyond the recognizable, but by choosing the correct sonification method. The work attempts to display technical data,

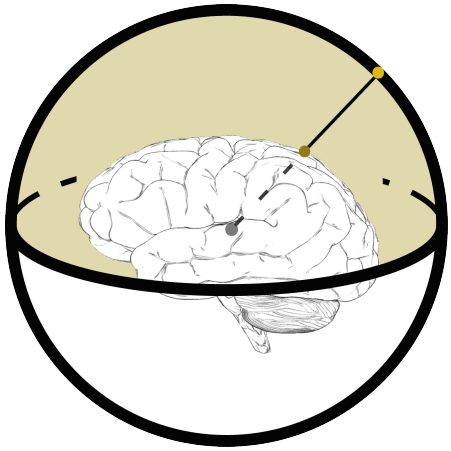


Figure 1: 3D projection of the features onto a virtual hemisphere. The grey dot represents both the center of the brain and the center of the sphere. The dark yellow dot represents the feature to be projected into the light yellow dot.

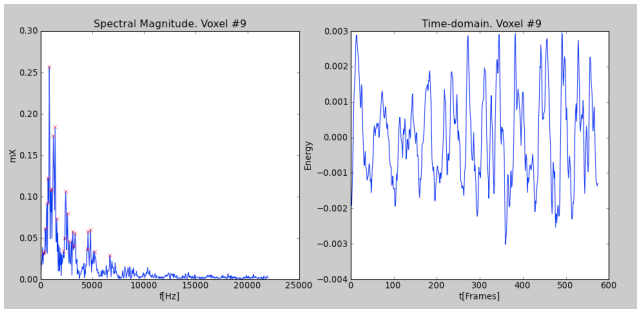


Figure 2: Time– Frequency representation from a voxel. The right graphic corresponds to the time domain, while the left graphic represents the magnitude of the voxel's spectral analysis.

while conveying feelings and make the experience enjoyable, both in terms of sonification and of spatial composition.

The first assumption is to consider each voxel as an audio sample to derive a base material to be later sonified. The approach taken can be somewhat compared to methods used in the spectral school [41]. Each voxel measurement contained around 500 samples. This is sufficient to extract a frequency analysis of the respective voxel. In the time domain, we normalize the samples and extract their mean. Afterwards proceed with the spectral analysis of the signal, and determine the most relevant frequencies. An example of a single voxel can be seen in Figure 2.

The most relevant frequencies are then mapped to the corresponding pitches. This results numerous scales and chords, each relating to different groups of the brain. Using these scales as compositional models, we can then score instrumental passages, which are performed and recorded as the base audio. Each passage contains a chord in its temporal center that represents the respective voxel as a whole. Compositionally, we then interpolate between the voxels.

This first step demonstrates the highest level of abstraction.

While each scale and chord represents a single voxel essentially, the amount of transformation done is beyond the recognizable. But the intention here was not to sonify the brain but to derive sound material that is only *based* on its data. The tonal composition itself is coarse, because, as it will be described further down, the spatial composition is able to distort the original sound to such degrees that it may claim the complete work for itself. Nevertheless, we retain the freedom to steer this basic material to our liking and create a well sounding instrumental composition.

4.1. Rapid panning modulation synthesis

On a less abstract representation of the brain lies the spatial composition. While the tonal composition was a necessity, the spatial considerations are the main focus of this work. The aesthetic followed here is similar to that of Hagan [35], in the sense that it creates a single engulfing sound. But while Hagan works with textures so dense that she describes a parallax between perceiving a single grain of sound and the complete, surrounding agglomerate as a single entity, we chose to pan our base material at speeds beyond the perception of motion.

In fact, the method described here is similar to how Stockhausen describes a technique used in *Sirius* (1975-77): "Sirius is based entirely on a new concept of spatial movement. The sound moves so fast in rotations and slopes and all sorts of spatial movements that it seems to stand still, but it vibrates. It is an entirely different kind of sound experience, because you are no longer aware of speakers, of sources of sound – the sound is everywhere, it is within you. When you move your head even the slightest bit, it changes color, because different distances occur between the sound sources"[2]

This above quote describes the sensation of the rapid panning modulation synthesis quite well. Once beyond the point that the motion of the source can be detected, the sound becomes static while still maintaining pulsating sensation. It becomes a single sound that is inherently spatial, meaning that the sound *becomes the space* as you cannot localize it any more even though is obviously present. Therefore, this work is not concerned about spatializing sounds in the traditional sense, it is about creating and working with *spatial sounds*. Furthermore, due to the omnipresence of the sound, the movement of the audience member inside lets him experience the sonorities differently. Hence, exploring both the auditory space and sound becomes one.

For *Sirius*, Stockhausen used a directional, rotary speaker to create this type of movement. Instead, for this work, we created a Max/MSP patch that is able to pan between an arbitrary amount of virtual loudspeakers on a sphere. This means that the actual sound source, as seen from the spatialization technology, is not moved, but the sound is sent to different virtual sources based on equal distance panning. This is done in both azimuth and elevation and the source signal can be panned by two modulation signals simultaneously in any direction.

Also, once the panning speed exceeds $\sim 20Hz$ in either direction sound synthesis is applied. The resulting effect is similar to amplitude modulation, but demonstrates significant differences. For one, the source sound theoretically is present in one to two speakers at a time. This means that the synthesis is a bit more complex and rich in high frequencies. More significantly, though, the rapid panned synthesis is *highly spatial*, meaning, it can not live without its space. If all virtual sources are moved into one another the synthesis is removed and the original sound surfaces.

[2] Stockhausen, as quoted in [26]

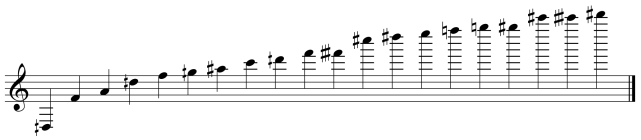


Figure 3: A set of notes extracted from the analysis in Figure 2.

4.2. Connecting the brain

Using a virtual loudspeaker setup instead of sending audio to the speakers directly brings many advantages. For one, the software that drives the artwork is independent of respective speaker setup on site. Furthermore, virtual speakers can be created at will and each speaker introduces a point of entry for further synthesis methods.

Having the complex spatial sound, we decided to introduce the voxels into the spatialization process by connecting their energy values directly with a filter. As the voxels were grouped into 50 regions on the half sphere, we used 50 virtual speakers, each with an individual processing unit. The voxel energy information was sent between two computers over the Opend Sound Control protocol, being normalized between $[0, 1]$. The information could then easily be rescaled to a respective center frequency. Additionally, the degree of change can be measured within a window and scaled to a meaningful Q-value.

The result is a colored, fully engulfing and pulsating sound. As the center frequencies of the many filters follow the energy values of each respective voxel region, the coloring of the whole construct is in constant shift, following the progression of the brain itself. Surprisingly, the sound was mostly uniform at first. But individual voxels started to break away from the large background, creating new auditory streams. Their position in space plays a key role. While a small number of voxels break away on their own, they create choreographies together, working with one another, against each other, from different points on the compass or next to each other, exchanging timbres and fusing to a single auditory stream.

5. CONCLUSIONS AND FUTURE WORK

As seen in the paper we have implemented a three dimensional sonification of fMRI brain data with aesthetic intentions. The brain data was filtered and projected onto a sphere. The sonification process was mainly carried out in two steps: first we derived pitched material from a quite abstract spectral analysis of each voxel, composing a base material from this pool of information. We then spatialized this data with a rapid panning technique creating a fully engulfing sound to represent the base material of the brain. Individual filters for each voxel then directly represents the activity and invites the visitor to explore this world with his own spatial hearing.

Visitors have reported a soothing, almost hypnotizing affection. Most were aesthetically pleased. The reduction of pitched material and other traditional musical parameters shifted the focus of the spatial interplay of each voxel successfully and made the composition/installation a true immersive experience.

For future work, we intend to investigate the interplay between different individuals whose fMRI data was recorded. Also, there are many points at which the sonification may tap in using different, higher level features. For example, as it can be seen in Figure 2, there is a clear low frequency oscillation in the time domain representation of a voxels energy development, which could be separated from the smaller fluctuations when subtracted, and used as two separate sonification methods. Also, we would like to group different meaningful regions of the brain, such as cerebellum, together, which could prove useful for macro-parameters or similar.

6. REFERENCES

[1] I. Xenakis, *Formalized music : thought and mathematics in composition*. Stuyvesant, NY: Pendragon Press, 1992.

[2] A. Schoon and F. Dombois, "Sonification in music," in *Proc. of the 15th Int. Conf. on Auditory Display*, Copenhagen, Denkmark, May 2009, pp. 76–78.

[3] D. A. Maluf and P. B. Tran, "Sensing super-position: Human sensing beyond the visual spectrum," in *IEEE International Conference on Information Reuse and Integration*, 2007, pp. 595–602.

[4] T. Hermann, M. Hansen, and H. Ritter, "Combining Visual and Auditory Data Exploration for finding structure in high-dimensional data," *Multimedia Systems*, 2001.

[5] S. Barrass and G. Kramer, "Using sonification," *Multimedia Systems*, vol. 7, no. 1, pp. 23–31, 1999. [Online]. Available: http://dx.doi.org/10.1007/s005300050108

[6] K. Vogt, T. Bovermann, P. Huber, and A. de Campo, "Exploration of 4d-data spaces. sonification in lattice qcd," in *International Conference on Auditory Display*, Paris, France, June 2008.

[7] F. Grond, S. Janssen, S. Schirmer, and T. Hermann, "Browsing rna structures by interactive sonification," in *Proceedings of ISon 2010, 3rd Interactive Sonication Workshop*, Copenhagen, Denkmark, 2010.

[8] D. O. R. Michael Winters, Andrew Blaikie, "Simulating the Electroweak Phase Transition: Sonification of Bubble Nucleation," in *Proceedings of the 17th International Conference on Auditory Display (ICAD2011)*, Budapest, Hungary, 2011.

[9] N. Bearman, "Using sound to represent uncertainty in future climate projections for the United Kingdom," in *Proceedings of the 17th International Conference on Auditory Display (ICAD2011)*, Budapest, Hungary, 2011.

[10] R. L. Alexander, J. A. Gilbert, M. Simoni, T. H. Zurbuchen, A. Arbor, and D. A. Roberts, "Audification as a Diagnostic Tool for Exploratory Heliospheric Data Analysis," in *Proceedings of the 17th International Conference on Auditory Display (ICAD2011)*, Budapest, Hungary, 2011, pp. 24–27.

[11] E. D. Adrian and B. H. C. Matthews, "The Berger Rhythm: potential changes from the occipital lobes in man," *Brain*, vol. 57, pp. 355–384, 1934.

[12] T. Hermann, P. Meinicke, H. Bekel, H. Ritter, H. M. Mueller, and S. Weiss, "Sonifications for eeg data analysis," in *Proceedings of the 8th International Conference on Auditory Display (ICAD2002)*, R. Nakatsu and H. Kawahara, Eds., Kyoto, Japan, 2002. [Online]. Available: Proceedings/2002/HermannMeinicke2002.pdf

[13] A. Hunt and T. Hermann, "THE IMPORTANCE OF INTERACTION IN SONIFICATION," in *Proceedings of the 10th Meeting of the International Conference on Auditory Display*, Sydney, Australia, 2004.

[14] T. Hermann, G. Baier, U. Stephani, and H. Ritter, "Vocal sonification of pathologic eeg features," in *Proceedings of the 12th International Conference on Auditory Display (ICAD2006)*, London, UK, 2006, pp. 158–163. [Online]. Available: Proceedings/2006/HermannBaier2006.pdf

[15] G. Baier, T. Hermann, and U. Stephani, "Multi-channel sonification of human eeg," in *Proceedings of the 13th International Conference on Auditory Display (ICAD2007)*, G. P. Scavone, Ed. Montreal, Canada: Schulich School of Music, McGill University, 2007, pp. 491–496. [Online]. Available: Proceedings/2007/BaierHermann2007.pdf

[16] I. Gomez and R. Ramirez, "A data sonification approach to cognitive state identification," in *Proceedings of the 17th International Conference on Auditory Display (ICAD2011)*, Budapest, Hungary, 2011.

[17] D. Wu, C. Li, Y. Yin, C. Zhou, and D. Yao, "Music composition from the brain signal: representing the mental state by music." *Computational intelligence and neuroscience*, 2010. [Online]. Available: http://dx.doi.org/10.1155/2010/267671

[18] S. Barrass, M. Whitelaw, and F. Bailes, "Listening to the Mind Listening: An Analysis of Sonification Reviews, Designs and Correspondences," *Leonardo Music Journal*, vol. -, pp. 13–19, 2006. [Online]. Available: http://www.mitpressjournals.org/doi/abs/10.1162/lmj.2006.16.13

[19] R. Melara and T. OBrien, "Interaction between synesthetically corresponding dimensions," *Journal of Experimental Psychology: General*, vol. 116, pp. 323–336, 1987.

[20] A. Einbond and D. Schwarz, "Spatializing timbre with corpus-based concatenative synthesis," in *Proceedings of the International Computer Music Conference*, New York, NY USA, 2010.

[21] D. Smalley, "Space-form and the acousmatic image," *Organised Sound*, vol. 12, no. 01, pp. 35–58, 2007.

[22] M. A. Harley, "Space and spatialization in contemporary music: History and analysis, ideas and implementations," Ph.D. dissertation, McGill University, PDF Reprint under Maja Trochimzyck, Moonrise Press, Los Angeles, California, 2011, 1994.

[23] R. Normandeau, "Timbre Spatialisation: The medium is the space," *Organised Sound*, vol. 14, no. 03, pp. 277–285, 2009.

[24] G. Theile, "Wave field synthesis – a promising spatial audio rendering concept," in *Proc. of the 7th Int. Conference on Digital Audio Effects*, Neaples, Italy, 2004, pp. 125–132.

[25] V. Pulkki, "Spatial sound generation and perception by amplitude panning techniques," PhD Thesis, Helsinki University of Technology, 2001.

[26] E. Bates, "The Composition and Performance of Spatial Music," Ph.D. dissertation, Trinity College Dublin, 2009.

[27] B. Zelli, "Reale und virtuelle Räume in der Computermusik," Ph.D. dissertation, Technische Universität Berlin, 2001.

[28] R. Zvonar, "A history of spatial music," *eContact!*, vol. 7, no. 4, 2006. [Online]. Available: http://cec.concordia.ca/econtact/Multichannel/spatial_music.html

[29] J. W. Solomon, "Spatialization in music: The analysis and interpretation of spatial gestures," Ph.D. dissertation, University of Georgia, May 2007.

[30] M. Trochimczyk, "From circles to nets: On the signification of spatial sound imagery in new music," *Computer Music Journal*, vol. 25, no. 4, pp. 39–56, 2001.

[31] H. Brant, "The uses of antiphonal distribution and polyphony of tempi in composing," *American Composer's Alliance Bulletin*, vol. 4, no. 3, pp. 13–15, 1955.

[32] V. Lombardo, A. Valle, J. Fitch, K. Tazelaar, and S. Weinzierl, "A Virtual-Reality Reconstruction of Poème Électronique Based on Philological Research," *Computer Music Journal*, vol. 33, no. 2, pp. 24–47, 2009.

[33] J. Smalley, "Gesang der Jünglinge: History and Analysis," 2000. [Online]. Available: http://www.music.columbia.edu/masterpieces/notes/stockhausen/GesangHistoryandAnalysis.pdf

[34] M. Fowler, "The Ephemeral Architecture of Stockhausen's Pole für 2," *Organised Sound*, vol. 15, no. 03, pp. 185–197, Oct. 2010.

[35] K. Hagan, "Textural Composition and its Space," in *Sound and Music Computing Conference: sound in space-space in sound*, Berlin, Germany, 2008.

[36] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA, USA: The MIT Press, 1990.

[37] B. Arons, "A Review of The Cocktail Party Effect," *Journal of the American Voice I/O Society*, vol. 35–50, no. July, pp. 701–705, 2001.

[38] M. Chion, "The three listening modes," in *Audio/Vision: Sound on Screen*. New York, NY, USA, availbale online: http://helios.hampshire.edu/∼hacu123/papers/chion.html: Columbia University Press, 1994.

[39] B. Blesser and L.-R. Salter, *Spaces Speak, are you listening?* Cambridge, MA, USA: MIT Press, 2007.

[40] J. a. Grahn and J. B. Rowe, "Feeling the beat: premotor and striatal interactions in musicians and nonmusicians during beat perception." *The Journal of neuroscience : the official journal of the Society for Neuroscience*, vol. 29, no. 23, pp. 7540–8, Jun. 2009. [Online]. Available: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2702750\&tool=pmcentrez\&rendertype=abstract

[41] G. Grisey, "Tempus ex machina: A composer's reflections on musical time," *Contemporary Music Review*, vol. 2, pp. 239–275, 1987.

"TRAINED EARS" AND "CORRELATION COEFFICIENTS": A SOCIAL SCIENCE PERSPECTIVE ON SONIFICATION

Alexandra Supper

Maastricht University,
Faculty of Arts and Social Sciences, Department
of Technology & Society Studies
P.O. Box 616, 6200 MD Maastricht,
the Netherlands
a.supper@maastrichtuniversity.nl

ABSTRACT

This paper presents a social science perspective on the field of sonification research. Adopting a perspective informed by constructivist science and technology studies (STS), the paper begins by arguing why sonification is an interesting case study to reconsider the role of sensory representation in scientific practice, and in particular the creation of credibility in science. It then focuses on a debate in which the meaning of objectivity is negotiated within the sonification community, showing that different notions of objectivity and scientific quality co-exist within the community, which are linked to different research questions being asked with the sonifications, different users that are envisaged for the sonifications, and different disciplinary backgrounds of the sonification researchers.

1. INTRODUCTION

The vast majority of papers about sonification and auditory display to date have been written by authors who are active in the creation or evaluation of auditory displays themselves. Only rarely has sonification attracted the attention of scholars in the social sciences or humanities [1], [2], [3]. To be sure, a number of scholars from within the sonification community have started to give thought to the historical underpinnings [4], [5], the philosophical implications [6] or the sociological context [7] of sonification. However, these contributions have generally approached their topics from within the logic of sonification research; that is, they have taken up themes that emerged within sonification work and used concepts or knowledge from these social science or humanities disciplines to think them through.

In this paper, I want to begin by taking the reverse approach. By adopting a perspective in the social sciences and humanities – and specifically, science and technology studies (STS) – I want to first ask *not* how history, philosophy or sociology can help us to understand sonification, but rather, how sonification can help to deepen our historical, philosophical and sociological understanding of how science works. I do so not because I expect the ICAD community to be full of closet social scientists, but rather, because I hope that beginning in such a way will allow members of the community to comprehend why a social scientist such as myself might be interested in sonification in the first place, as well as to understand the perspective I have chosen in my research. By zooming in on debates about conceptions of objectivity within the sonification community, I then want to suggest how such STS research might also help the community to understand some of its own struggles. I do so by outlining two different perspectives on objectivity which coexist within the ICAD community, which I refer to as the 'trained ears' and the 'correlation coefficients' approaches.

2. RESEARCH CONTEXT

The research described in this paper is part of a larger project on the sonification of scientific data, adopting a perspective informed by science and technology studies (STS) [8]. Like sonification, STS is often described as an interdisciplinary field or an emerging discipline, encompassing perspectives from fields such as the sociology, history and philosophy of science and technology [9], [10]. The common denominator of STS work is an interest in the interactions between science, technology and society. Notably, these interactions cannot be reduced to talking about "societal impacts" of science and technology, but also involve the many ways in which the development of science and technology is itself shaped by societal and cultural aspects.

This project is dedicated to the study of the ICAD community (as the institutionalized embodiment of sonification) as well as examples of sonification from the world of electronic music and science popularization. It tries to understand the popular appeal and fascination of sonification, as well as its scientific legitimacy. In doing so, it adopts a constructivist perspective, assuming that what is or is not accepted as legitimate and credible science is not a matter of course, nor can it be determined by hard-and-fast universal criteria that distinguish science from non-science, but is in fact the product of an ongoing negotiation process [11]. Accordingly, my research attempts to trace how the scientific legitimacy of sonification is negotiated by various actors inside and outside of the sonification community.

Methodologically, the research described here is based on a qualitative analysis of a number of different empirical

sources: semi-structured qualitative interviews with practitioners of sonification; participant observation research at sonification-related conferences, workshops, talks and concerts; and primary texts, such as conference proceedings, journal articles and dissertations.

3. SONIFICATION AND THE HIERARCHY OF THE SENSES

Philosophers, anthropologists and historians of science and culture have agreed for a long time that there exists such a thing as a "hierarchy of the senses", and that the sense of vision possesses an established seat at the top of this hierarchy [12]. Sight has been argued to be strongly linked to rationality, detachment and science, in contrast to the supposedly more emotional and subjective sense of hearing [12], [13]. However, detailed empirical research in STS and the history of science has recently complicated and nuanced this picture somewhat, calling into question the inevitability of the development towards a visual culture of science. Instead, these researchers have shown that other senses also play a role in scientific practice [14], [15], as well as that the scientific status of vision, too, has been frequently contested [16], [17].

This is not to say, however, that the sense of vision is unimportant in science – indeed, visual elements are ubiquitous in scientific practice [18]. It means that what kind of sensory representation or evidence will be accepted as scientifically credible is not a matter of course; rather than taking for granted that the sense of vision will always dominate, it is up to the STS researcher to analyze, based on detailed empirical studies, what is, or is not, accepted as part of credible and legitimate scientific research in certain contexts. Rather than assuming that vision will always be associated with detachment and rationality, and that sound will always create subjective and emotional experiences, it becomes crucial to study the historical and cultural processes in which precisely these connotations are created, strengthened, challenged or negated.

Sonification is a particularly apt case for such a study precisely because it questions the traditional hierarchy of the senses; that is, it calls into question the commonplace assumption that the only 'proper' way of dealing with scientific data is to visualize them. The rules and conventions that might otherwise remain invisible because they are taken for granted become explicit and observable when an alternative method for the representation of scientific data is proposed. Understanding sonification, and especially its scientific legitimacy and the strategies used to establish its credibility, therefore adds further nuances to the understanding of what is accepted as scientifically legitimate in different contexts, and how this sense of legitimacy is created.

4. A SHORT HISTORY OF OBJECTIVITY

The question of what does and does not count as a scientifically legitimate representation of data is closely intertwined with notions of scientific objectivity. Indeed, the terms "objective" and "scientific" are often used synonymously. However, STS researchers – most notably, Lorraine Daston and Peter Galison [19] – have argued that objectivity has not always been considered a defining ingredient of science, and indeed, that the concept of objectivity itself has a history: the term has been used to signify different characteristics in different contexts and settings. Instead of trying to identify whether particular scientific practices are or are not objective with the help of a checklist, these authors have argued that objectivity itself is a historically constructed and mutable concept; a concept that cannot be nailed down to one fixed meaning but is negotiated in relation to specific practices and representations.

On the basis of an analysis of images in scientific atlases, Daston and Galison trace the historical construction of scientific objectivity, showing how the "epistemic values" of science have changed over the centuries [19]. They focus on three such epistemic values in particular: truth-to-nature, mechanical objectivity, and trained judgment. The ideal of truth-to-nature guided science until the 19th century. In this regime of representation, scientific atlas-markers sought to abstract from the individual idiosyncrasies and imperfections that exist in nature, in favor of a higher plane of perfection and a depiction of ideal types. As an emblematic example of truth-to-nature, Daston and Galison discuss an image in a botanical atlas, in which "the underlying type of the plant species, rather than any individual specimen" [19] was depicted.

In the late 19th century, truth-to-nature gradually started giving way to the ideal of mechanical objectivity. With the emergence of mechanical objectivity, the presence of a human observer became problematic and the depiction of idealized archetypes was very much frowned upon; instead, the actual specimens, with all their peculiarities and irregularities, now moved to the front-stage. Letting nature speak for itself, with the help of machines that were supposedly uncontaminated by human influences, was now the goal of scientific depiction. To illustrate the representational practices of mechanical objectivity, Daston and Galison reprint an image of a snowflake, which "is shown with all its peculiarities and asymmetries" [19].

In the 20th century, yet another epistemic value emerged and took its place alongside truth-to-nature and mechanical objectivity: trained judgment. If truth-to-nature sought to distill the idiosyncrasies of scientific specimens into an idealized representation, and mechanical objectivity tried to do away with any kind of human intervention and interpretation in order to let nature speak for itself, then the emergence of trained judgment marked a point where human intervention and interpretation became permissible again. However, trained judgment was not oriented towards the creation of idealized images, but rather the detection of patterns and structures in large amounts of data. With the help of trained eyes and other tacit skills, scientific specialists learned to distinguish between relevant and irrelevant characteristics in the data, and no longer shied away from enhancing visualizations to better display the attributes of interest. As a characteristic image for the practice of trained judgment, Daston and Galison discuss a visualization of the magnetic field of the sun, in which "the output of sophisticated equipment [was mixed] with a

'subjective' smoothing of the data" to remove instrumental artifacts [19]. According to Daston and Galison, the emergence of this regime of representation was strongly linked to the existence of a new generation of professionally trained scientists brimming with self-confidence in their scientific judgment.

Daston and Galison's work, however, is based entirely on a study of visual representations of science, specifically the graphic illustrations used in scientific atlases; they do not consider that these, or other, epistemic values of science might also be linked to different forms of representation, such as auditory displays. In this paper, I want to extend their work on the historical constructions of objectivity into the domain of auditory representations. In particular, the concept of trained judgment will also come in useful for understanding sonification.

5. THE CONTESTED OBJECTIVITY OF SONIFICATION

The objectivity of auditory displays of scientific data is frequently contested. Many ICAD researchers have shared anecdotes about peer reviewers or potential collaborators who have dismissed the possibility of sonifying data out of hand, without even seriously considering its potential advantages. Interestingly, however, sonification is contested even among some of those scientists who do in fact make use of it.

That is to say, there are a number of scientists who work with sonification, while at the same time denying its scientific legitimacy. For instance, some asteroseismologists tend to play audifications of stellar oscillations while giving popular talks in order to convey something about their research to lay audiences, and yet insist that this has nothing to do with their actual research. They argue that these are just helpful gimmicks in the process of science popularization, but that sound plays no role in their analyses.[1] Sonification is thus used, but simultaneously disavowed as a serious scientific component.

By framing sonification in this way, these scientists do not call into question traditional ideas about vision as the only sense that is compatible with rationality, objectivity and serious scientific research; in fact, they reinforce them by making a clear distinction between proper science (characterized by numbers and images) on the one hand, and popularized science (which may also involve sound) on the other hand. And indeed, the fact that they frame sonification in this way shows just how deeply engrained these ideas about the hierarchy of the senses and about the subjectivity of sound have become in the minds of many scientists.

However, other framings of sonification and its objectivity also exist, and it is these that I want to turn to in the remainder of this paper. Particularly within the ICAD community, debates are taking place about how to position sonification in order to establish its scientific legitimacy and objectivity. These concerns are related to ambitions of formalizing the community and "to encourage increased standards and increased quality of the papers"[2] at the annual conferences, as Bruce Walker, then president of ICAD, put it.

6. "CORRELATION COEFFICIENTS"

This, of course, raises the question of what the criteria for a good publication would be. For Walker, who was trained as a psychologist and computer scientist, the question seems relatively clear-cut. He makes a distinction between contributions that contain research components and those that are just "doing show and tell" [3]. In this distinction, contributions with a research component are marked by their theoretical contextualization, and especially by efforts of evaluation. Another long-standing ICAD member talks about the importance of evaluations and user tests in similar terms:

> You need some way to measure what you actually achieve when you're using sonification. It's not enough that you say this, listen, this really sounds better than yesterday. That's not the result. But if you can show that when you have 10 people doing this task they do things 10% better when they're using the auditory display than when they're not using the auditory display – that's a result.[4]

In this quote, scientific quality is clearly defined in terms of quantification: the qualities of a good sonification can be demonstrated with the help of hard numbers and backed up by correlation coefficients and other measures of statistical significance. Essentially, this addresses the objection that sonification cannot be objective because the sense of hearing is subjective and because it cannot be guaranteed that information is indeed accurately picked up by listening. It does so by quantifying what the average listener actually hears in a sonification, or how he or she works with this information. For the sake of brevity, I have referred to this way of thinking about the scientific quality as the 'correlation coefficients' approach.

And indeed, it is not the only conceptualization of scientific quality that exists within the ICAD community; some members are very critical of the insistence on user-tests, claiming that there exists a tendency of "evaluating oneself to death."[5] This becomes most explicit in an anecdote about an argument related to peer review decisions at a previous ICAD conference:

> Many of the best sonification examples were curated out, peer reviewed away. (…) There is a central stream and poster sessions, and [many] good things were sent into the poster sessions. Because [the reviewers] had abstruse ideas about evaluability and intersubjectivity. So they said, if somebody makes a sound and did not make a series of user tests with 17 (...) test persons, then we cannot accept this, because that's not scientific. It's as if you would not have a graph printed if someone cannot prove that he let 17

[1] Interviews with the asteroseismologists Conny Aerts (March 2009) and Donald Kurtz (November 2009).

[2] Interview with Bruce Walker (June 2009).

[3] Interview with Bruce Walker (June 2009).

[4] Interview with Matti Gröhn (July 2009).

[5] Interview with Florian Grond (June 2008).

> people look at the graph to make sure they can see something in the graph. That is, I think, that's absurd.[1]

Several ICAD members have expressed criticism of such an insistence on user-testing, arguing that sonifications that are novel and innovative can be very valuable and inspiring to the community even if they do not come with an evaluation.[2] Besides, it has been suggested that user-tests often consist of rather trivial tasks focusing on the qualities that are easiest to measure rather than those that are in fact most relevant for potential users of the sonification.[3]

7. "TRAINED EARS"

However, the critics of (mandatory) user-tests in sonification research do not advocate that sonification should refrain from making claims to scientificity and real research. Rather, they offer an alternative conception of the objectivity and scientificity of sonification, one which is not necessarily linked to quantitative evaluation. In reference to Daston and Galison's 'trained judgment' [19], discussed in section 4, I have called this paradigm the 'trained ears' one.

Analogous to what Daston and Galison refer to as trained judgment, the supporters of a trained ears approach defend the acceptability of a certain amount of subjective decisions, provided they are made with the help of the trained ears of scientific experts. Subjectivity is explicitly embraced here – but paradoxically, without giving up claim to objectivity altogether:

> The reader may have the impression that such sonifications are so strongly tuned to the subjective preferences of the user that they may not be particularly 'objective' to communicate structural features in the data. However, sonification is actually always the result of strongly subjective tuning of parameters. Furthermore, each mapping is equally valid as true representation of the data. Only the combination of different (sonic) 'views' may yield a more 'objective' overall impression of structures in the data. [20]

In this quote, Thomas Hermann and his co-authors discuss a combination of different 'sonic views'. By changing certain parameters and listening to different versions of the same dataset, different acoustic perspectives are provided. It is through listening to many, many different sonifications (and possibly glancing at visualizations at the same time) that the researcher fully begins to understand the overall structure and patterns that are contained in the data. In contrast to the user-testing paradigm, emphasis is put not on the creation of intersubjectivity by having different test persons listen to the same sonification, but rather on having the same person listen to different displays of the same dataset. Every single one of these displays may be subjective, but a trained listener is able to detect patterns in the data.

The term 'sonic views' is also interesting because it explicitly likens sound to vision. Indeed, many proponents of the paradigm of trained ears invoke the authority of visualization. Often with reference to the fact that subjective decisions are widely accepted in data visualization and not usually second-guessed through perception tests,[4] it is argued that sonification should not be required to have to prove its usefulness time and time again. Instead, sonification experts should have the self-confidence to trust that sonification can in principle provide trustworthy displays of scientific data; and in order to really make the most out of a specific sonification application, the proponents of this approach argue, experts in the concrete research subject should be closely involved with the making of the sonifications. Once the expert opinions of domain scientists have been involved in this way, the argument goes, "there will be good reason to trust not only the judgment of a visualization expert about a picture, but also a judgment of a sonification expert about a sound" [21]. Since the number of relevant scientific experts is often too small to allow for a quantitative evaluation in the proper sense, quantitative evaluation may not necessarily be appropriate in such cases [22], [23]. Besides, especially when it comes to developing sonifications that are meant to be used in exploratory research, the kind of well-defined tasks that can be tested for easily may not really be of interest at all; it is difficult to devise quantitative tests for complex, unpredictable and long-term research questions.[5]

8. INTERDISCIPLINARY FRICTIONS

In the previous two sections, I have indicated that different conceptions of objectivity and scientific quality exist within the ICAD community. So far, I have only hinted at how these differences can be explained. Is this a case of irrationally feuding camps or of haphazardly differing epistemological tastes? In this section, I want to show that this is not the case; rather, both positions can be explained as outcomes of different research questions being asked, different users envisaged for the sonification applications, and different disciplinary backgrounds.

Research Questions

The ICAD community is connected by a shared interest in the usage of sound to convey information, but the underlying research interests that bring different individuals into the community may vary considerably. While some in the field are primarily interested in aesthetic issues, others might emphasize informational requirements; while some are interested in investigating general capabilities of the human auditory system and exploiting this knowledge by designing applications according to these features, others might be more interested in using sonification as a tool for the analysis of

[1] Interview with Florian Dombois (February 2008). The quote has been translated from German by the author.

[2] Interview with Thomas Hermann (October 2009).

[3] Interviews with Alberto de Campo (October 2009) and Florian Grond (June 2008).

[4] Interviews with Florian Dombois (February 2008) and Florian Grond (June 2008).

[5] Interview with Alberto de Campo (October 2009).

complex data; while some use sonification to build audio interfaces for particular devices, others concentrate their efforts on exploring particular datasets via sound.

These different research interests also entail different requirements of empirical verification, and therefore rub off on what the researchers consider appropriate standards of valid scientific research. For instance, for someone who is primarily interested in auditory perception research, it is essential to find out general features of the human auditory system, and it is therefore plausible to involve relatively large numbers of subjects when putting a sonification to the test, as the perception of human subjects is at the very core of the research interest. On the other hand, for someone who is primarily interested in the development and implementation of new techniques for data mining and display, such user tests may be a means to an end or a nice extra, but they are not an essential component of the research itself. Just how important it is considered to involve large numbers of listeners in testing sonifications is therefore very much related to the precise research questions being tackled through sonification.

Sonification Users

Another difference exists in the people that the sonification researchers have in mind as (potential) users for their applications. Some ICAD members explicitly follow a 'universal design' approach.[1] This term specifically refers to the equal inclusion of people with and without disabilities – in the case of sonification, in particular the inclusion of blind as well as sighted users – but at the same time also suggests broader implications: the user being targeted is, for all intents and purposes, 'everybody'. Now, if 'everybody' is intended as a user of a sonification, it also makes sense to try to involve 'everybody' in the testing of a sonification. While actually involving everybody is impossible in practice, large-scale user tests are built on the idea of providing an approximation of this: if not everybody, then at least the average user.

On the other hand, there also exists a tradition of developing sonifications specifically for expert users, such as scientists working on a specific line of research. As mentioned above, these cases may be less amenable to quantitative testing, as the targeted user group may be too small to allow statistically significant quantitative evaluations. What is more, the ideal image of this type of sonification research is often built upon an intensive and sustained collaboration between sonification researchers and scientific specialists.[2] In those cases, evaluation may happen in much more informal and incremental forms in the course of the collaboration, and a formal evaluation may be deemed unnecessary or an unwanted burden for the scientific specialist who already invests a lot of time and effort into an unusual type of research with uncertain results. For exploratory scientific research, the best empirical evidence of the usefulness of a sonification may not be an auditory perception test, anyway, but rather the discovery of a new scientific insight by means of listening, which could then be substantiated by other means and lead to theoretical advances.

Disciplinary Backgrounds

The different views on the necessity of user-tests are also related to different disciplinary orientations that co-exist within the ICAD community. Of course, this aspect is not independent of those discussed above; for instance, the type of research questions being asked are very much related to disciplinary perspectives. But different disciplines not only bring different research questions to the table; they also have their own, not necessarily compatible, quality standards and conceptions of objectivity [24], [25]. And yet, these standards are often taken as self-evident and universal.

For instance, the requirement of user-testing is often taken as inevitable due to one's disciplinary training; as one ICAD member reflects, "I'd been kind of trained in the way, from the viewpoint that you always have to do an evaluation, otherwise you can't state whether you've given a contribution or not."[3] From within a particular disciplinary perspective, a particular type of testing may seem like the most natural and unavoidable thing in the world, yet discipline-specific standards should not be mistaken for universal ones. The user-testing paradigm, for instance, is in fact strongly related to a psychological tradition of quantitative experimentation. Not only is it connected to one particular scientific discipline, rather than to general scientific principles, but even within the discipline of psychology, the development of such an experimental tradition was a contingent rather than an inevitable one.

As historians of psychology have shown, psychologists have drawn upon strategies of standardized testing and quantitative measurement in an effort to demarcate their discipline from the muddy waters of the humanities and common sense. Instead, a close alignment with the natural sciences was sought by emphasizing methodological similarities. In short, then, the strong reliance on tests and experiments was a particular historical strategy to establish the cultural authority of psychology by emphasizing its affinity with already established natural scientific disciplines [26], [27].

My claim here is not that user-testing is a phenomenon exclusive to psychology; indeed, the practice has taken strong roots in other disciplines too, including some – such as human-computer interaction – that have a strong foothold within sonification. Nor do I want to call into question the value of such tests. I do, however, want to point out that they have roots in a very specific historical and cultural context, and should not be mistaken for inevitable and universal ingredients of scientific work.

It has become clear in this paper that such an approach is not shared by everyone within the ICAD community. Specifically, I have sketched out an alternative to the 'correlation coefficients' approach, which I have referred to as 'trained ears'. It is more difficult to associate this approach

[1] Interviews with Bruce Walker (June 2009) and Stephen Brewster (November 2010).

[2] Interviews with Thomas Hermann (February 2008), Florian Grond (June 2008) and Alberto de Campo (October 2009).

[3] Interview with Paul Vickers (January 2011).

with any particular type of evaluation practice; after all, one of its tenets holds that a systematic evaluation might not be necessary as the involved researchers should trust their own expert judgement. Where close collaboration between sonification researchers and domain specialists is sought, the quality standards of the involved domain science (be that neurology, seismology, sociology or chemistry) might be as relevant as whatever standards the ICAD community can come up with; this is especially true when a publication in an academic journal in the data domain is aspired. It is therefore no surprise that some researchers within ICAD are more reluctant about favouring the quality standards of any particular scientific field.

This does not mean that the debates about the need for evaluations within the sonification community run neatly along disciplinary lines, nor that there exist two full-fledged and clearly defined competing camps. However, the difficulties of finding agreement on the appropriate quality standards is rooted in a scientific culture in which different research interests and disciplinary backgrounds meet, and in which no consensus has been established about what the standards for good scientific work could be. This is not unusual; sociological studies have shown that agreement on quality standards in interdisciplinary fields is often difficult, because different disciplines come with their own ideas and standards of quality. In fact, this is particularly true for fields that also involve input from outside the confines of academic science [24], which is the case for sonification with its strong connections into art and design. It is no surprise, then, that the method of user-testing as evidence for the scientific quality of sonification is controversially discussed within the ICAD community.

9. CONCLUSIONS

In this paper, I have highlighted one angle from which sonification can be a fruitful object of studies for STS. As I have argued, sonification can be interesting to the STS researcher because it opens up new perspectives on the types of representations that are considered permissible in scientific practice. The case of sonification shows that it is not self-evident that scientific analyses are made and scientific results presented only in a visual form, as sound can also be used to represent scientific data. At the same time, however, it also shows that scientific conventions favor visual rather than auditory displays. Auditory displays tend to be marginalized in scientific practice, and those researchers who do want to make use of them adopt different strategies to counter this marginalization. It has not been my goal to predict the success of these strategies, but rather to examine the logic according to which they operate.

While my starting point was a perspective asking how sonification can be of interest to STS, rather than how STS can be useful for sonification, my paper can also contribute to the discussions and self-reflection of the sonification community. Specifically, by analyzing the debates about objectivity within the ICAD community in terms of two conflicting paradigms – which I have termed 'trained ears' and 'correlation coefficients' – I have elucidated some of the positions that are taken within the community. I have not only sketched out these two paradigms, but shown that each of them can be explained in terms of different research questions, different envisaged users, and different disciplinary backgrounds. Most importantly, I have shown that each of these traditions of thinking about the objectivity of sonification research has a history and has to be understood in a particular historical and cultural context. It is in this way, I believe, that STS research of the kind I have undertaken here can be of interest to the ICAD community, as it shows how positions taken in such debates are shaped by sociological and historical factors.

Above, I have referred to the two paradigms as "conflicting", but this is not meant to imply that they are incompatible by definition; indeed, there is nothing in these two positions that would preclude them from co-existing peacefully within the same community. To do so, however, both would have to be accepted as valid and equitable scientific approaches by everyone in the field. In the current constellation, the two are often pitted against in each other in the search for appropriate quality standards for the field as a whole.

This desire to agree on quality standards itself has to be understood in a particular historical and cultural context. It forms part of an ongoing process of professionalization, in which the community strives for clearer professional standards and markers of quality. This process is so urgent precisely because what is at stake is not just the acceptance or rejection of specific papers at the conference; what is at stake is how the field presents itself to the outside (scientific) world, and the standing of sonification research as a whole. The fact that it is difficult to agree on shared standards of quality has much to do with the interdisciplinary nature of the sonification field. There seems to be much awareness within the community of the fact that evaluation criteria differ between scientific, engineering and artistic projects, but little explicit attention has been paid to the fact that even the criteria of different scientific fields may differ; let alone to how these differences specifically play out in debates within the community or in peer review decisions. In this paper, I hope to have elucidated some of these differences and thus to contribute to the community's process of self-reflection. More than anything, though, I look forward to discussing my findings with the ICAD community.

10. REFERENCES

[1] J. Sterne and M. Akiyama, "The recording that never wanted to be heard, and other stories of sonification," *The Oxford Handbook of Sound Studies*, T. Pinch and K. Bijsterveld, eds., pp. 544-560, Oxford, UK: Oxford University Press, 2012.

[2] V. Straebel, "The sonification metaphor in instrumental music and sonification's romantic implications," in *Proc. of the 16th Int. Conf. on Auditory Display (ICAD)*, Washington, DC, 2010, pp. 287-294.

[3] A. Supper, "The search for the "killer application": Drawing the boundaries around the sonification of scientific data," *The Oxford Handbook of Sound Studies*,

T. Pinch and K. Bijsterveld, eds., pp. 249-270, Oxford, UK: Oxford University Press, 2012.

[4] G. Kramer, "An introduction to auditory display," *Auditory Display. Sonification, Audification and Auditory Interfaces*, G. Kramer, ed., pp. 1-77, Reading: Addison-Wesley Publishing Company, 1994.

[5] F. Dombois, "The muscle telephone: The undiscovered start of audification in the 1870s," *Sounds of Science – Schall im Labor (1800-1930)*, J. Kursell, ed., pp. 41-45, Berlin, Germany: Max Planck Institute for the History of Science.

[6] D. Worrall, *Sonification and Information: Concepts, Instruments and Techniques*, PhD thesis, Canberra, 2009.

[7] A. de Campo, C. Dayé, C. Frauenberger, K. Vogt, A. Wallisch and G. Eckel, "Sonification as an Interdisciplinary Working Process," in *Proc. of the 12th Int. Conf. on Auditory Display (ICAD)*, London, UK, 2006, pp. 28-35.

[8] A. Supper, *Lobbying for the Ear: The Public Fascination with and Academic Legitimacy of the Sonification of Scientific Data*, PhD thesis, Maastricht, 2012.

[9] S. Jasanoff, G. E. Markle, J. C. Petersen and T. Pinch, eds., *The Handbook of Science and Technology Studies: Revised Edition*, Thousand Oaks: SAGE Publications, 1995.

[10] E. J. Hackett, O. Amsterdamska, M. Lynch and J. Wajcman, eds., *The Handbook of Science and Technology Studies, Third Edition*, pp. 165-180, Cambridge: The MIT Press, 2008.

[11] T. F. Gieryn, *Cultural Boundaries of Science: Credibility on the Line*, Chicago, IL: The University of Chicago Press, 1999.

[12] R. Jütte, *A History of the Senses: From Antiquity to Cyberspace*, Cambridge, UK: Polity Press, 2005.

[13] C. Classen, *Worlds of Sense: Exploring the Senses in History and Across Cultures*, London, UK: Routledge, 1993.

[14] C. C. M. Mody, "The sounds of science: Listening to laboratory practice," *Science, Technology, & Human Values*, vol. 30, no. 2, pp. 175-198, spring 2005.

[15] R.V. Burri, C. Schubert and J. Strübing, "Introduction: The five senses of science. Making sense of senses," *Science, Technology & Innovation Studies*, vol. 7, no. 1, pp. 3-7, May 2011.

[16] P. Galison, *Image & Logic: A Material Culture of Microphysics*, Chicago, IL: The University of Chicago Press, 1997.

[17] A. Beaulieu, "Images are not the (only) truth: brain mapping, visual knowledge, and iconoclasm," *Science, Technology, & Human Values*, vol. 27, no. 1, pp. 175-198, winter 2002.

[18] R. V. Burri and J. Dumit, "Social Studies of Scientific Imaging and Visualization," *The Handbook of Science and Technology Studies: Third Edition*, E. J. Hackett, O. Amsterdamska, M. Lynch and J. Wajcman, eds., pp. 297-317, Cambridge: The MIT Press, 2008.

[19] L. Daston, and P. Galison, *Objectivity*, New York: Zone Books, 2007, pp.20-21.

[20] T. Hermann, K. Bunte, and H. Ritter, "Relevance-Based Interactive Optimization of Sonification," in *Proc. of the 13th Int. Conf. on Auditory Display*, Montréal, Canada, 2007, pp. 461-467.

[21] F. Dombois, O. Brodwolf, O. Friedli et al., "SONIFYER. A Concept, a Software, a Platform," in *Proc. of the 14th Int. Conf. on Auditory Display (ICAD)*, Paris, France, 2008.

[22] A. de Campo, R. Hoeldrich, G. Eckel et al., "New Sonification Tools for EEG Data Screening and Monitoring," in *Proc. of the 13th Int. Conf. on Auditory Display*, Montréal, Canada, 2007, pp. 536-542.

[23] K. Vogt, F. Plessas, A. de Campo et al., "Sonification of Spin Models: Listening to Phase Transitions in the Ising and Pott Model," in *Proc. of the 13th Int. Conf. on Auditory Display*, Montréal, Canada, 2007, pp. 258-265.

[24] K. Huutoniemi, "Evaluating Interdisciplinary Research," *The Oxford Handbook of Interdisciplinarity*, R. Frodemann, J. T. Klein and C. Mitcham, eds., pp. 309-320, Oxford: Oxford University Press, 2010.

[25] M. Lamont, *How Professors Think: Inside the Curious World of Academic Judgment*. Cambridge, MA: Harvard University Press, 2009.

[26] M. G. Ash, "Historicizing mind science: Discourse, practice, subjectivity," *Science in Context*, vol. 5, no. 2, pp. 193-207, autumn 1992.

[27] T. Dehue, "Deception, efficiency, and random groups: Psychology and the Gradual Origination of the Random Group Design," *Isis*, vol. 88, no. 4, pp. 653-673, december 1997.

NON-SPEECH AUDIO-SEMIOTICS
A REVIEW AND REVISION OF AUDITORY ICON AND EARCON THEORY

David Oswald

Design in Business Communication Management
HTW Berlin University of Applied Science,
10313 Berlin, Germany
oswald@htw-berlin.de

ABSTRACT

The aim of this paper is to develop a theory and taxonomy of auditory signs, based on semiotics. For more than two decades, the discourse on non-speech audio interfaces has been dominated by a dichotomy between auditory icons, which are based on everyday hearing, and earcons, which are based on musical hearing. The corresponding theory behind these concepts has to be revised for several reasons. First, the authors of these theories partly use semiotic concepts and terminology, but not always in a correct way. Second, the classification of auditory icons as "iconic", and earcons as "abstract" is too simple and based on the questionable premise that everyday sounds are per se iconic and musical motives are per se abstract and symbolic. Third, this widespread idea ignores the crucial role of the user in the process of perception. In addition, the users' perception of visual and auditory signs in computer interfaces is fundamentally different today, from how it was in the early years of graphical user interfaces — the time when the first auditory interfaces and the corresponding theories were developed.

1. INTRODUCTION

Computers operate with several layers of symbolic code ranging from binary machine code to high level programming languages. Therefore, strictly speaking, all signs in human computer interfaces are symbolic — at least on a technical level. Iconic signs have been introduced to human interfaces by a metaphoric transfer from the actual world to the computer model world. Visual icons have served as a model for both auditory icons and earcons [1], [2]. The related theory construction drew parallels between auditory and visual icons. Literature on both, auditory icons and earcons, has employed semiotic terms and definitions, but in some cases in a rather unorthodox way. The most common fallacies are the confusion over indexical and iconic signs, thus confusing *causality* with *similarity* [3], and the notion of earcons being purely conventional and symbolic [1].

In order to outline a semiotics-based theory of non-speech audio in human computer interfaces, the first necessary step is to correct these misbelieves. Not as an end in itself — a revised semiotic theory of auditory signs will also shed a different light on stereotype attributions concerning advantages and disadvantages of auditory icon and earcon use. It can be expected, that a better understanding of the semiotic processes will improve decision-making during the design process.

In a second step, the theory needs to be amended with respect to today's users who have grown up with digital media, the so-called *digital native*. The concepts of auditory icons and earcons were developed in the 1980s — at a time when graphical user interfaces and the desktop metaphor where still new and unfamiliar to the users. Today, many users have internalised the model world of the graphical interface to an extent that makes menus, icons, and windows actually feel "natural". This habituation effect strongly influences also the perception of auditory signs, and hence, changes the semantic relation between the auditory sign and its meaning.

After a brief introduction to some basic terms of semiotic theory, these two steps of review and revision will be made. Based on semiotic definitions, a taxonomy of auditory signs in human machine interfaces will be suggested.

2. RELATED WORK

Semiotics for non-speech audio has been adressed systematically only in recent years. Pirhonen et. al. [4] and a related article of Murphy et. al. [5] have rightly adressed the fact that a sign's interpretation is influenced by its semiotic context (syntagma). However, they adhere to the distinction betweeen real-world auditory icons and earcons that are "symbolic in nature". Petocz et.al. [6] have clearly described the listener's essential role in the sign process. Nevertheless, their re-interpretation of auditory icons as "conventional indicators" can be questioned. Last, Nam and Kim [7] provide a (too) simple one-to-one mapping of sign classes to auditory cues. Whereas they use Peirce's refined "ten principal classes of signs" on the semiotic part of the equation, they use only a rather undifferentiated classification of auditory signals.

3. SEMIOTICS

Semiotics, the study of signs and sign processes, is rooted in philosophy and linguistics. Due to the modern semiotics' tradition of more than a century, the various semiotic schools and their respective terminology cannot be discussed here in detail. However, in order to discuss a semiotic theory of auditory signs, it is necessary to introduce to a minimum of semiotic terminology beforehand. In this article, the semiotic terminology will follow that of Charles Sanders Peirce, who introduced the triadic concept of the sign, which emphasises the role of the perceiving person in the sign process [8].

3.1. The three aspects of the sign

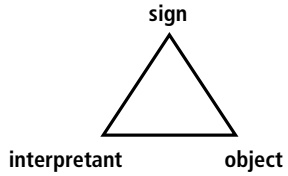


Figure 1: The three aspects of a sign, following Peirce [8].

1. Sign: the sign-carrier, the perceptible signal.
2. Object: the thing or the concept the sign refers to.
3. Interpretant: the interpretation in the mind of the perceiver.

It is seems somewhat confusing that one of the sign's parts is again called the "sign". In Peirce's terminology, it denotes the physically existing sign, which can be auditory, visual, haptic or olfactory. Some scholars refer to it as the "sign-carrier", "sign-vehicle", or the "signal". Eventually, mostly the more simple term "sign" is used.

In addition, the term "object" might be misleading. The object can be a physical object or thing, like a car or a trashcan, but it does not have to be physical. The object can also be an abstract concept like "democracy", or an action like "erase" [9].

Finally, the "interpretant" should not be confused with the interpreter, i.e. the interpreting person. It is rather the interpreter's mental conception of the sign's meaning. In other words, like the sign-carrier, the interpretant is a representation of the object. But whereas the sign exists physically in an auditory, visual, haptic or olfactory form, the interpretant exists "in one's head only". [8]:

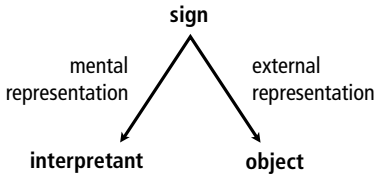


Figure 2: The sign refers to an external object and evokes a mental representation. Illustration by the author.

3.2. The three dimensions of semiosis

The sign process (semiosis) is subdivided into three dimensions that describe the relations between sign, object and interpretant [10]:

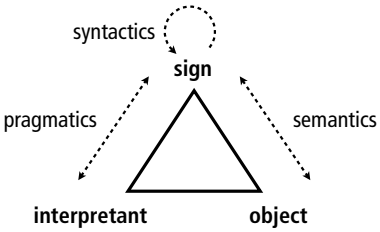


Figure 3: The dimensions of the sign process. Own illustration, partly based on Morris [11].

1. Syntactics: The relation between sign and other signs, rules for the formal structure of signs.
2. Semantics: The relation between sign and its object, its meaning.
3. Pragmatics: The relation between the sign and its interpretant, the effect the sign has on the perceiver.

3.3. The three types of relation between sign and object

Semantics are not only about the meaning of signs, but also about the principles behind the construction and encoding of their meaning. Semiotic theory differs between three types of signs, based on distinct relations between the sign and the referred object.

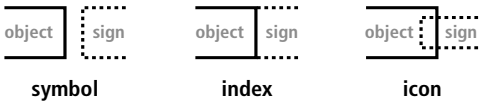


Figure 4: Schemes of the relations between signs and their object. Own illustration following Bense [12].

1. Symbol: based on *convention*, no factual link between sign and object
2. Index: based on *causality*, physical link between sign and object
3. Icon: based on *similarity* between sign and object

4. TYPOLOGY OF VISUAL AND AUDITORY SIGNS

The scientific discourse in the auditory display community has been utilizing some semiotic concepts and terminology, but — as will be discussed in chapter 3.1. — not in a consequent or consistent way. On the other hand, the semiotics community has hardly discussed the domain of non-speech audio.

Morris was the first to systematically apply semiotics to the visual domain [13] and to teach semiotics in a design context [14]. Today, semiotics is an integral part of the curricula of numerous graphic design study programs, but in auditory communications semiotics remain regrettably unutilized.

This blind spot of the semiotic discourse has its origin in the discipline's strong tradition in linguistics. Even in Nöth's extensive "handbook of semiotics" [15] only a small chapter on semiotics of music can be found, but the term "sound" is simply non-existent in the subject index. In musicology, there is also no great tradition in semantic analysis of music. The meaning of music, in the sense that it refers to extra-musical phenomena, is not in the focus of traditional art music theory. In most cases, musical analysis is mainly based on the syntactical and self-referential inner structures of music. Exceptions to this are the semiotic driven works of Tarasti [16], Nattiez [17], and Cummings [18], and Clarke's approach to musical meaning based on ecological perception [19]. In contrast to everyday sounds, music does not have an unambiguous meaning. If a piece of music has extra-musical meaning, it is often based on a complex, multi-layered, and interwoven symbolic (cultural) coding [16]. Hence, music is a form of communication with a great power of evoking associations and moods, but it is usually not used in a strictly functional context, that is to communicate well-defined

meanings effectively without ambiguity. However, below these multiple cultural layers there are also musical universals, which are independent of cultural context. For instance, the sense for tempo, and what is considered fast or slow, is similar across all cultural backgrounds. Musical universals can be used to design music-based signs that are *not* arbitrary and symbolic, and therefore are as easy to learn as natural everyday sounds. This will be discussed further in chapter 3.2.

4.1. Index

The most frequently used example for an indexical sign is smoke as a sign for fire. Smoke indicates a fire, and it does so by merely pointing to it, without being similar to the fire and without cultural conventions behind it [8]. The index sign is linked to its object simply by the laws of nature — it is a symptom. The auditory index sign for "fire" would be the fire's typical crackling sound. The fire physically causes this sound, it is the auditory effect of physical and chemical processes that we call "fire". The index sign "crackling" and its object "fire" are linked so closely, that one could argue that "smoke" and "crackling" are both integral parts of the perceiver's conception of "fire". Everyday listening is mainly based on these indexical sign processes. Gaver also points to the direct and effortless perception of physical everyday sounds:

Our normal mode of hearing is to listen to sounds to identify the events that cause them. From this perspective, sound provides information about materials interacting at a location in an environment. [2]

4.2. Icon

Most definitions of the iconic sign use the term "similar" to characterize it. Thereafter, the icon is based on a similarity between the sign and what it stands for [8]. In order to be more precise, Morris circumscribes the concept of similarity with "shared attributes between sign and object" [11]. The iconic principle of similarity is widely used in visual communications. For instance, a silhouette drawing of an animal on a traffic sign becomes understandable by the depiction's similarity to the animal. Sign and object share some attributes of shape. Iconic auditory signs in this sense would be sounds that *sound* similar like other sounds. Foley artists often use iconic sounds, for instance when using coconut shells to imitate horses, or when using a snare drum as an exaggerated illustration of a punch in the face.

A recording of a sound is, when played back, an icon for the original sound. Digital photo cameras use pre-recorded mechanical shutter sounds to indicate an otherwise silent digital process. When originally produced by a mechanical camera, this sound is a physically caused index sound for the shutter release. Everyone who is familiar with analogue photo cameras understands this indication intuitively. Therefore, when a digital camera reproduces a shutter sound, the imitated sound is interpreted due to its similarity with the original sound. It is an auditory icon. But what about younger users, who are not familiar with vintage photo gear? For them the meaning of the same sound is pure convention — a symbol. [20]

4.3. Symbol

The well-known error beep is a typical example for an auditory symbol. Symbols are based on mere convention, neither laws of nature nor perceivable similarity link a symbol to its meaning [8]. The sign's shape or sound has no factual connection with what it refers to, which is why the symbolic sign often is referred to as being *arbitrary*. The traditional error beep is in fact arbitrary, in the sense that its timbre, pitch and duration do not contribute anything to its meaning. A higher or lower pitch or a different waveform would do the same job just as well. Pure waveforms, like sine waves, lack physical indexical meaning because they are hardly heard in everyday interaction with the environment. They can only obtain a meaning by declaration and convention [20]. But what about using real world sounds, like frog's croak or glass bottle sounds, as a sign for a computer error? In relation to their actual meaning, these are just as arbitrary as a sine wave. Originally, they are index sounds, which indicate for instance the presence of a frog. Transferred to a different context the indexical meaning retreats to the background and gets overlaid by the new symbolic meaning. It is only a matter of repetition and training until the second meaning becomes dominant [6].

Multilayered meanings are not restricted to digital technology, for instance the sound of a church bell is initially only an indexical sign for a clapper hitting a metal bell-shaped vessel. Still, the predominant meaning of this sound is the appeal to attend church service, or the profane indication of the current time. Both of the latter codes work on a symbolic level, based on initially arbitrary cultural conventions — other cultures use different sounds for these purposes. Even this arbitrary coding can be perfectly internalized in a way that it will be understood just as fast and intuitively as natural indexical sounds [21].

4.4. Iconicity

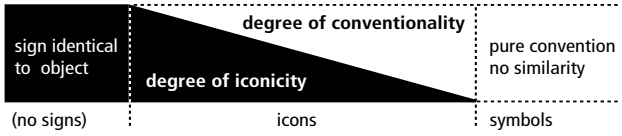


Figure 5: Gradual transition of icon to symbol, from high iconicity to high conventionality [22], [23].

In order to discuss the typology of auditory signs further, it is necessary to have a closer look at *similarity*. In the visual domain, it seems to be obvious when a sign is similar to its object. The silhouette drawing of a cow on a traffic sign is said to be similar to a real, living cow — at least in some aspects. Here similarity is based on proportional scaling, reduction to two dimensions, elimination of materiality and colour, and reduction of details in shape.

However, similarity is not restricted to analogue transfers like scaling or reduction of detail. A merely diagrammatic similarity is also considered to be iconic [8]. Even if a subway map is not drawn to scale, or a circuit diagram does not represent the spatial arrangement on the circuit board, both are still iconic representations based on structural similarity.

In order to describe different levels of similarity between sign and object Morris introduced the term "iconicity" [22]. In this sense, the attraction of Madame Tussauds' wax figures is based on a very high iconicity, whereas a subway map is based

on low iconicity. The upper end of the iconicity scale is delimited by a sign that is identical to its object, and therefore would not be a sign anymore. Below the lower end of the iconicity scale's is a sign that has no (more) similarity with its object — a symbol [23]. This delimitation is not defined by objective properties of the sign, but solely by the perception of the interpreter. If a low level similarity is recognised or not, depends strongly on the perceiver's previous knowledge, cultural background and frequency of use [24].

The concept of iconicity as a degree of similarity is easily understood when dealing with visual icons. Similarity of auditory icons is harder to define, since natural sounds are signs for *events*, they are *time-based*, whereas visual icons represent *things*, they are *spatial*. In the following chapters, the question of iconicity of auditory signs and to what they actually are similar will be addressed.

4.5. Using index, icon and symbol

Taking a superficial view, an index sign seems to be the most intuitive sign to be understood, because it is "natural". The second choice would be the icon, because it bears the potential to be understood by resemblance. The symbol would be coming in last, as "arbitrary" usually is considered almost synonymous with "inapprehensible". While it is undoubted that different sign-object relations exist, it must also be clear that in terms of understandability the different types of sign are only good for a head start effect. All described advantages can and will be overridden by the effects of repetitive use. Moreover, in fact index signs are not more intuitive because they are "natural" — they have become intuitive only because we have been exposed to them for a longer time.

The given description of the three types of signs has been simplified in order to be clear and concise. In fact, also the interpretation of indices and icons are to a certain extent subject to cultural differences and context. For a discussion on the cultural influence on the perception of "direct physical experience" (i.e. index signs), see Lakoff and Johnson [25]. For a discussion on the conventionality of icons and on perception of similarities as a cultural technique, see Eco [26].

5. AUDITORY ICON AND EARCON THEORY

The terms "auditory icon" and "earcon" have been coined in the 1980s, when the discipline of auditory computer interface design emerged. In the early years, the discourse has been dominated by a methodological debate about which of the two concepts is more effective and easier to learn. Today *both* are standards in auditory display design. Both concepts have constituted the (still improvable) auditory environment of today's computer users. Browsing for instance sound folders of Apple's OS and Microsoft Windows, auditory icons and musical earcons can be found in peaceful coexistence. This is also reflected in scientific discourse: A cumulative word count through the ICAD proceedings of the past three years shows 490 hits for the term "earcon" and 356 hits for "auditory icon", with an average of six occurrences (!) per paper.

In order to reconceive auditory icon and earcon theory, it is necessary to once again have a look at classic publications, which coined and imprinted these terms, since some debatable attributions that originate from these early papers keep being repeated until today. The most problematic stereotypes in this context is the notion that auditory icons are per se iconic, and that earcons are generally abstract, i.e. symbolic.

5.1. Are auditory icons really iconic?

It is needless to say that Bill Gaver's work on auditory icons [2], [3] has been seminal for auditory display design. In his dissertation, Gaver transferred Gibson's approach of ecological perception [27] from the visual to the auditory domain [28]. He analyses how information can be obtained from everyday sound and discriminates it strictly from musical hearing. Due to their intuitive understanding, Gaver recommends the use of everyday sounds for auditory interface design. In his argumentation, he refers to Peircian semiotics, but obviously confuses index and icon when he claims that "iconic mappings are based on physical causation" and "its characteristics are causally related to the things it represents" [3]. This is true for indices, but not for icons, which are not based on causality but on similarity. This flaw has been noted before by Petocz et. al. who then conclude that auditory icons in fact are auditory *indices* [6]. However, the matter is even more complex.

As we have seen in chapter 2.1, everyday sounds are indexical. But what happens when these sounds are being detached from the event of their physical causation? A recorded and played back sound could be described as an index for a past event. Even with the best high fidelity equipment, the recorded sound will not be exactly the same like the original sound. Hence, a played back sound is, due to its similarity to the original sound, only a representation of the original sound and thereby also a representation of the original sound's meaning. Gaver's argument that auditory icons are iconic because they are based on physical causation is not correct. However, only the explanation was wrong. They are iconic, because they have been *copied and imitated*, as we have seen in the example of the camera shutter sound in chapter 2.2.

Admittedly, the camera example is different to most computer interface scenarios. In the shutter sound example, the original context and the new application context reside in the very same domain. In contrast, computer interfaces do not have mechanical predecessors that could serve as a source of physical sounds and established listening habits. Everything in a graphical user interface is based on metaphors. Using a trashcan to delete data on a computer seems almost natural today, but of course, it is based on a conceptual analogy between throwing away waste in real live, and marking hard disc space as unused. In real live, the accompanying sound when trashing something is an integral part of the perceptional pattern of “trashing”. A visually similar representation of a trashcan, a similar interaction and a similar sound create a holistic multisensual analogy in the computer model world — and *iconic* sign-object relations in all of the three aspects: visually, auditory and interaction-wise. Such coherence in all aspects of a conceptional model is rare, because many processes in computers do not have an analogue equivalent in real live.

Some of Gaver's auditory icons in the "Sonic Finder" [3] were an extension of the visual desktop metaphor with what can be called an auditory "carpenter metaphor": Applications sounded like metal, like tools do. Files and folders — the material to be worked with — sounded like wood. Are these metaphoric signs also iconic? In what sense is a wooden sound similar to a digital file? What attributes do they share? These attributions do not seem to be built on similarity in the usual sense. Although, taking a look into the classic definition of "meta-

phor", we again come upon the concept of similarity. Following Aristotle, a metaphor can be a transfer based on the principles of *analogy*, which is in turn based on similarity or comparability [29].

However, a metaphor does not have to build on an already existing similarity. The similarity is rather created by the introduction of the metaphor [30]. Aristotle already pointed out that coming up with a good metaphor "implies an intuitive perception of the similarity in dissimilars" [29]. Thus, the sounds of files and folders in Gaver's "Sonic Finder" are iconic indeed. They are based on a conceptional similarity between metal/wood, tool/material and application/file. A similarity that came into life by the metaphorical transfer introduced by Gaver.

Already Peirce described three kinds of iconic similarity: A picture sharing basic qualities with its object, a diagram displaying relations only, and a metaphor where the similarity refers to yet another sign [8]. In Morris' terms, a metaphor is an iconic sign with low iconicity.

So far, we only considered Gaver's metaphoric mapping of file type to material and timbre. Based on this metaphor he also proposed mapping the file size to pitch, so that — analogous to real life experience — big objects would produce low pitch sounds and small objects would produce high pitch sounds. This mapping has been coded into the sound-producing algorithm, with file size as the parameter that determines the sound's pitch [31]. Thus, file size and pitch correlate in a fully predictable and reproducible way. This suggests a *causal* relation — which is untypical for icons, but constitutive of *index* signs. Here, causality is not based on the laws of physics, but rather on man-made rules written into a software algorithm. In this sense, parametrised sounds act on an indexical level.

In conclusion, signs that are based on everyday sounds are not necessarily auditory *icons*. When there is not even a metaphorical similarity between auditory signs and their meaning, for instance when a frog's croak is used as an alert sound, then even a natural everyday sound is simply arbitrary and symbolic. More complex are parametrised auditory icons that have at least two semantic layers in which meaning is encoded concurrently; the metaphoric icon with low-iconicity where timbre denotes the file type, and a second indexical layer where for instance pitch has an algorithmic, causal relation with file size. If these layers are both perceived equally, or if one layer becomes dominant, is eventually depending on the listener.

5.2. Are earcons really abstract, i.e. symbolic?

The counterpart to Gaver's first publications on auditory icons was the paper "Earcons and Icons: Their Structure and Common Design Principles" by Blattner et al. [1]. In this paper the authors coined the term "earcons" and defined them as auditory signs based on musical principles — short micro-compositions of only a few notes length.

Even if very short, earcons do share their design parameters with music: tempo and rhythm, melodic gestalt, timbre, dynamics, harmonics. Nevertheless, the authors mostly address parallels between earcons and visual icons as well as methods to create modular earcon families. Surprisingly, a discussion of how musical parameters can be used to convey meaning — the semantic impact of musical parameters — has been left aside completely. For instance, tempo and melodic gestalt obviously evoke strong associations, which can and should be utilized when designing earcons. Instead, the authors are content with the notion that earcons, in contrast to auditory icons, are abstract and symbolic and therefore simply have to be learned [1]. For Blattner et. al. the only way to facilitate earcon learning is a systematic and hierarchical earcon design. To speak in semiotic terms, it is a completely syntactic approach, ignoring semantic aspects of music. This compares to describing principles for writing readable text, while only focusing on grammar and spelling.

The concept of earcons as basically arbitrary compositions leads to a problematic negligence towards the actual composition of the earcon. Compared to everyday sounds, which are always indices for their causing events, it is much more difficult to describe the meaning of music. Music is widely considered being self-referential, bare of any extra-musical meaning. This may be true in some cases for "pure" art music. Programme music and especially functional music, like film music, show impressively how music is able to transport not only moods, but also information that can hardly be transmitted visually, such as the existence of monsters under a bed, or a protagonist's hidden feelings. These denotations are in most cases coded in multiple layers of cultural conventions, but there are also aspects in music that are directly understood, independent of musical training and across cultural differences. These so-called musical universals are based on biological and physiological structures [32], or rooted in human perception [19]. The sense of tempo correlates perfectly with both heartbeat and walking; 120 beats per minute are considered a fast tempo in music, a fast heartbeat rate, and also a fast walking pace. Universal music related patterns are also found across different spoken languages. An excited speaker will speak louder and faster, in a higher pitch, using greater intervals — features that are also used to describe excitement in musical theory [32].

The terms "high" and "low" pitch suggest a correlation between pitches and physical space. Indeed, most people associate a change in pitch with motion in an imaginary space. If this association is based on a physiological effect, is still being debated [33]. However, ecological approaches to the perception of musical meaning regard the association of motion as directly rooted in human perception [19]. Even if the effect was only culturally acquired, it is anchored into our listening habits so deeply that it is impossible to ignore when designing earcons. In Microsoft Windows, simple two-tone motives indicate when hardware has been added or removed. In fact, there is no objective reason for assigning an ascending interval to "adding" and an descending interval to "removing", but to match "in" with "up" and "out" with "down" fits listening habits and therefore feels intuitively right.

Longer motives can create a more complex contour or *gestalt*. Gestalt theory, originally developed in cognitive psychology in order to explain phenomena in visual perception, has also been applied to describe the perception of melodic patterns [34]. Tempo and melodic gestalt are just two examples to illustrate the non- arbitrariness of earcons. Rhythm, dynamics, and timbre also carry connotations that can and should be utilized in earcon design. The concept of gestalt had already been addressed during the very first ICAD conference in 1992 [35]. However, it did not lead to doubts about the concept of earcons as completely abstract and symbolic signs.

In conclusion, earcons can be completely arbitrary and symbolic, but they do not have to be. When a simple synthetic beep represents a system error, the beep is an arbitrary symbol. More complex earcons can also be arbitrary, for instance when

the famous four-note motiv of Beethoven's 5th symphony would be used to indicate "added hardware". However, there is a plethora of associations evoked by musical universals that can be utilized in earcon design in order to serve a communicative goal. Already a sequence of only two tones produces a notion of tempo, a directed motion, and a melodic gestalt with qualities like fast or slow, flowing or hesitant, up or down, and calm or volatile. In this case, meaning is based on similarity between patterns of musical perception on one side, and analogous perceptional patterns of extra-musical phenomena on the other. A musical tempo may have similarity with familiar timing patterns of strolling, walking, or running. These similarities are mainly metaphorical, since they cross domains like pitch and physical space. In this case, attributes from an original domain (i.e. physical space) are used to denote attributes in an alien domain (i.e. pitch). In consequence, well-designed earcons that build on musical universals are in fact *iconic* signs with low iconicity, for they make use of metaphorical similarity.

6. SIGN METAMORPHOSIS

The relation between the sign and its object does not exist objectively. It is not a fixed property of the sign. Whether a sign is perceived as indexical, iconic or symbolic does not solely depend on the quality and the characteristics of the sign, in fact it depends on the sign process as a whole. In which way a sign is interpreted by a perceiver is strongly depending on their previous knowledge and the present context. The same sign may be understood on a similarity basis by one perceiver and simply by habit and convention by another. Still, in large groups of perceivers, there are predominant patterns of interpretation. However, these predominant patterns of interpretation may change over time. In his theory of sign metamorphosis, Keller has described the shifting semantic relations between signs and their objects, and the changing ways of how a perceiver derives meaning from a signal [24].

6.1. From index to icon

When an index is imitated, it becomes an icon. To illustrate this effect, Keller uses the example of a simulated yawn. A real yawn is an index for a shortage of oxygen. Like index signs in general, yawning is usually not used for intentional communication. However, a simulated yawn can serve as an effective iconic sign for letting someone know how bored the listeners are. It is understood because it is similar to the real yawn. The same rule applies to auditory signs. As seen in chapter 2.2, a camera shutter sound becomes an iconic sign by imitation — it is then interpreted by an associative inference, based on the similarity between the original and the recorded sound. [24]

6.2. From icon to symbol

Whereas an icon becomes meaningful by an association that is triggered by perceived similarity, a symbol obtains its meaning by conventions, i.e. written or unwritten rules. Keller points out that the associative way in which iconic similarity is interpreted, is a creative process without normativity. It is always possible that the interpreter has an association different from the intended goal. This procedure of association can be compared to solving riddles. Confronted with the same riddle for several times, one does not have to associate and guess anymore. Therefore, by repetitive use, an icon will not be interpreted by similarity any more but based on a habit, a rule. The similarity actually is still there, but now remains unnoticed. The similarity has become useless. In consequence, iconic signs that are used frequently over long periods of time will lose more and more of their iconicity by simplification and abstraction. A visual example is the metamorphosis of the iconic cipher III to the symbolic 3, which developed over the centuries by cursive handwriting and rotation by 90°. In everyday conception, the cipher 3 is a symbol for most people, until they learn about the relation to its iconic predecessor III and start to see the visual similarity. Then the cipher 3 has again become an icon — for just as long as the similarity remains conscious. [24]

6.3. From any sign to index

In Keller's linguistic perspective the described sign metamorphosis is a one-way street where signs start as indices or icons, and become symbols at the end [24]. This may be true for spoken language, but is not necessarily the case in digital interactive systems. When we interact with interfaces, we continuously interpret visual and auditory signs emitted by the system. These signs follow the logic that has been encoded into the system by the system's designer and are meant to be either indexical, iconic, or symbolic. Whereas repetitive use of iconic signs in the analogue world often leads to a symbolification of these signs, in digital interactive systems it leads to *indexicality*. Whatever sound is played back, when for instance a file is dragged to the trash, if it is only repeated often enough, it will become an *index* for the event of successfully putting a file into the trashcan. This can work even with the most arbitrary auditory cue; it will require only more repetition.

In the perception of a frequent computer user, it does not make a difference if a sound is determined by physical parameters when interacting with the real world, or if a sound is triggered by the user's interaction with the virtual world and determined by man-made algorithms. The only required condition that leads to an indexical sensation is *perceived* causality. When I *always* hear the same sound when trashing something, and when I *never* hear it when I missed the trash, then the sound becomes quickly an indicator for trashing — independent of the sound's features and qualities. In the user's perception, his or her activity in the computer model world *causes* this sound.

6.4. Polysemy

Usually the term polysemy is used to describe ambiguous or multiple meanings of a sign. Thereafter a "beetle" can denote either an animal or a car, and in spoken language, it could also denominate John, Paul, George, or Ringo. An outline drawing of a man may represent a man — or a bathroom in a different context. In the latter example not only the meaning changes, but also the sign-object relation. In the first case, the relation is iconic, for the drawing visually resembles a man. In the second case, the relation is symbolic, because the depicted manikin does not share any visual attributes with the signified bathroom.

So there is obviously also a second meaning of polysemy, which does not deal with multiple meanings but with multiple types of sign-object relation. In Keller's theory of sign metamorphosis we saw that these relations change, from index to icon by imitation, and from icon to symbol by frequent use.

	indices	icons		symbols			
everyday sounds »auditory icons«	parameterised sounds	played back index sound orignal and new context within the same domain	played back index sound crossing related domains			arbitrary index sounds from arbitrary domains	interface sound
	causal relation by algorithm	strong similarity of physical sound attributes	metaphoric similarity an index sound's original meaning is a symbol for the intended meaning			pure convention, no factual relation, original meaning irrelevant/misleading	sign-object -relation
	pitch ≙ file size	real trashcan's *crash* sound ≙ virtual trashcan's sound	baby's *scream* ≙ error alert			frog *croak* ≙ error alert	example
musical motives »earcons«	any sound can become indexical by habituation		motives based on musical universals	motives based on cultural connotation of music	single abstract tones	arbitrary musical motives	interface sound
	perceived causal relation by algorithm		metaphoric similarity, patterns in musical perception similar to analogous perceptive patterns	multilayered symbolic meanings, reduced arbitrariness by tradition and cultural imprinting	pure convention no factual relation, no similarity, no causation	pure convention no factual relation, musical meaning irrelevant/misleading	sign-object -relation
			ascending interval ≙ add hardware	fanfare ≙ king ≙ glory ≙ success	simple *beep* ≙ error alert	5th symphony ≙ add hardware	example
			5th symphony ≙ error alert				

Figure 6: Proposed taxonomy of auditory signs.

It is important to note that these change processes do not proceed simultaneously or in a regulated way for all users [24]. Thus, a sign can be interpreted on a similarity basis, and *at the same time,* someone else may interpret it based on mere habit or convention. For the first interpreter it is an icon, whereas it is a symbol for the second. Still, in spite of their different ways of making sense, both interpreters can derive the very same meaning at the end. Concluding this chapter, we can say that it is hard, or almost impossible, to predict in which *way* a perceiver will interpret an auditory sign. Nevertheless, it is comforting to see that the intended meaning can still come across, even if in different ways.

7. INDEXICALITY: DIGITAL NAÏVES, DIGITAL IMMIGRANTS AND DIGITAL NATIVES

When Gaver published his first article on auditory icons in 1986 — only two years after the introduction of the Apple Macintosh — graphical user interfaces (GUI) where still new and unfamiliar to most computer users. Computer users who were confronted with iconic representations of files, folders, printers and trashcans on a computer screen, rightly conceived these icons as *representations* of something. Icons were perceived consciously as signs that stand for digital, symbolic, and invisible code. Users were very aware that the desktop metaphor is a *metaphor*, and that it was designed to facilitate learning to use a computer.

As explained in the previous chapter, signs change the way they are conceived for instance by frequent use. A similarity-based associative inference will be superseded by a rule-based inference or mere habit. Like this, icons become symbols. The initial iconic sign process is completely contingent upon the interpreter's ability to recognize or construct similarity [24]. In the 1980s these interpreters were inexperienced GUI users — digital naïves. In their everyday life, files and folders were physical objects made of paper and cardboard. In contrast, today's young adult users grew up with computers. These digital natives do not conceive the computer model world as a representation of an office [36]. Depending on their age, they probably did not even know paper files and cardboard folders before they encountered the corresponding representations on the screen. Therefore, for digital natives, these representations never were perceived as representations. Due to the lack of knowledge about the originally depicted objects, they were unable to construct any similarity. For them, representative "icons" were just arbitrary symbols. Hence, a semiotic explanation of the digital native phenomenon can be subsumed as *a sign metamorphosis taking a shortcut from symbols directly to indices*. The same effects apply to auditory signs. Neonates need some time to learn symbolic sounds like doorbells or police sirens. However, natural sounds also have to be learned in the first place. For instance discerning sounds of bouncing and breaking glass does not have to be easier than internalising a symbolic "beep" as an index for error. Hence, some of the advantages of everyday sounds are simply based on longer learning time.

In addition, *digital immigrants*, who did not grow up with digital technology but have adopted it, also develop indexical perception by continuous use. When the computer model world behaves consistently over long periods of time, when user interaction triggers predictable and reproducible feedback, then every user will soon internalise feedback signs and consider them as *indicative* for his or her actions. Like this, signs that once were consciously conceived as representations of something, become quasi-natural index signs. In the actual world, sounds are created by the natural law of physics. In the computer world sounds are caused by the laws of man-made algorithms. Once these algorithms are implemented, the sounds are determined by the user's interaction and the algorithms — the sounds can be internalized just like natural sounds.

8. CONCLUSION

Auditory icons and earcons cannot be attributed with fixed types of signs. In everyday life, natural sounds are indexical signs, based on causality. In an interface they can as well be iconic, or even completely symbolic and arbitrary. Earcons do have a tendency towards a conventional and symbolic coding of meaning. However, if composed attentively, they can also become iconic and intuitively meaningful (see figure 6).

The question of *how* a sound communicates its meaning, if a user makes sense of it by causality, similarity or convention, does have an effect on its learnability. However, since the type of sign (index, icon, or symbol) is *not* directly depending on the type of sound (everyday or musical), there cannot be a well-

defined rule of which type of sound is easier to learn. Learnability does not depend primarily on the *type* of sound, but rather on the distinct sound that is used, its characteristics, its sound design, composition, cultural connotation, or original context. Hence, especially in its early years, auditory interface discourse put too much emphasis on the types of sounds (everyday or musical). Over the discussion of principles, the concrete design of proposed and tested sounds has often been neglected — especially in the case of earcons and their proper composition.

Of course, a scientific community has to generate generalisable knowledge. However, generalisation should not lead to over-simplification. Labeling earcons with "abstract" and auditory icons with "concrete", and the deduced cliché that music-based earcons have to be learned, whereas everyday sounds are intuitively understood, are over-simplifications in that sense. In contrast to the sciences, design does not have to produce general truth. Design usually aims for specific solutions for specific users in a specific context. A rule in the manner of "use sound type A for purpose B with user C" is per se too simple to be valid. Of course there can be patterns or rules of thumb like this, but the fate of auditory signs is decided by the adequacy of their original context (everyday sound), their composition (musical sounds) and the specific sound design details.

Last, there is the decisive influence of the user: Habituation of individual users on one hand, and different ways of perceiving interfaces of the digital naïve, the digital immigrants, and the digital native on the other hand. A deeper understanding of these factors will allow for a better focus on the relevant issues of sound design for auditory display.

9. REFERENCES

[1] M. Blattner, D. Sumikava, R. Greenberg, "Earcons and Icons: Their Structure and Common Design Principles", in: *Human-Computer Interaction*, vol. 4, 1989, pp. 11-44.

[2] W. Gaver, "Auditory Icons: Using Sound in Computer Interfaces", in *Human-Computer Interaction*, vol. 2 no. 2, June 1986, pp. 167-177.

[3] W. Gaver, "The SonicFinder, An Interface That Uses Auditory Icons", *in Human Computer Interaction*, 4, 1989, pp. 67-94.

[4] A. Pirhonen, E. Murphy, G, McAllister, W. Yu "Non-speech sounds as elements of a use scenario: A semiotic perspective", *Proc. of ICAD 2006*, London, pp. 134-140.

[5] E. Murphy, A. Pirhonen, G. McAllister, W. Yu, "A Semiotic Approach to the Design of Non-speech Sounds", in *Proceedings of HAID 2006*, Glasgow, pp. 121-132.

[6] A. Petocz, P. Keller, C. Stevens, "Auditory warnings, signal-referent relations, and natural indicators: Re-thinking theory and application", *Journal of Experimental Psychology: Applied*, vol. 14, no. 2, Jun 2008, 165-178.

[7] Y. Nam, J. Kim, "A semiotic analysis of sounds in personal computers: Toward a semiotic model of human-computer interaction", in *Semiotica* 182, vol. 1 no. 4, pp. 269–284, 2010

[8] C. S. Peirce, "Logic as Semiotic: The Theory of Signs" in *The Philosophy of Peirce: Selected Writings*, London: Routledge & Kegan Paul, 1956, pp. 98-119.

[9] U. Eco, *A Theory of Semiotics*, Bloomington: Indiana University Press, 1979.

[10] C. Morris, "Foundations of the Theory of Signs", in *Writings on a General Theory of Signs*, The Hague: Mouton, 1971, pp. 17-71.

[11] C. Morris, "Esthetics and the Theory of Signs", in *The Journal of Unified Science (Erkenntnis)*, vol. 8, no. 1/3, The Hague, Holland, June 1939, pp. 131-150.

[12] M. Bense, *Zeichen und Design*, Baden-Baden: Agis, 1971.

[13] P. Betts, "New Bauhaus und School of Design, Chicago", in J. Fiedler, P. Feierabend, (ed.), *Bauhaus*, Cologne: Könemann, 1999, 66-73.

[14] C. Poisson, "De l'objet au sujet; pour une sémiotique du projet en design. Charles Morris et le New Bauhaus", in I. Rauch (ed.) *Semiotics around the world: synthesis in diversity*, Berlin: Mouton De Gruyter, 1997, pp. 859-862.

[15] W. Nöth, *Handbuch der Semiotik*, Stuttgart: Metzler, 2000.

[16] E. Tarasti, *Signs of Music – A Guide to Musical Semiotics*, Berlin: Mouton de Gruyter, 2002.

[17] J. Nattiez, *Music and Discourse – Toward a Semiology of Music*, Princeton University Press, 1990

[18] N. Cummings, *The Sonic Self: Musical Subjectivity and Signification*, Bloomington: Indiana Univ. Press, 2000.

[19] E. Clarke, *Ways of Listening – An Ecological Approach to the Perception of Musical Meaning*, New York: Oxford University Press, 2005.

[20] D. Oswald, *Sound in Computer Interfaces*, unpublished thesis, Cologne University of Applied Science, 1996.

[21] D. Oswald, "Semiotik auditiver Interfaces – Zur Geschichte von Gestaltung und Rezeption auditiver Zeichen in Computer-Interfaces" in A. Schoon, A. Volmar (eds.), *Das geschulte Ohr. Eine Kulturgeschichte der Sonifikation,* Bielefeld: Transcript, 2012, pp. 247–267.

[22] C. Morris, "Signs, Language, and Behavior", in *Writings on a General Theory of Signs*, The Hague: Mouton, 1971, pp. 79-397.

[23] C. Morris, *Signification and Significance: A Study of the Relation of Signs and Values*, Cambridge: MIT Press, 1964.

[24] R. Keller, *A Theory of Linguistic Signs*, New York: Oxford University Press, 1998.

[25] G. Lakoff, M. Johnson, *Metaphors We Live By*, Chicago: The University of Chicago Press, 1980.

[26] U. Eco, *Einführung in die Semiotik*, Munich: Fink, 1972.

[27] J. J. Gibson, *The ecological approach to visual perception*, Hillsdale: Lawrence Erlbaum Associates, 1979.

[28] W. Gaver, *Everyday Listening And Auditory Icons*, PhD Thesis, University of California, San Diego, 1988.

[29] Aristotle, *Poetics*

[30] M. Black, Metaphor, in: *Proceedings of the Aristotelian Society*, vol. 55, pp. 273–294, 1954

[31] W. Gaver, "Synthesizing Auditory Icons", in *Proceedings of INTERACT '93 and CHI '93*, New York, 1993.

[32] H. de la Motte-Haber, "Die Musik als Affektlaut", in *Welt auf tönernen Füßen*, Forum, vol. 2, Göttingen, 1994.

[33] P. Janata, "The Highs And Lows Of Being Tone Deaf", in *Nature Neuroscience* 10, 2007, pp. 810-812.

[34] M. Wertheimer, "Untersuchungen zur Lehre von der Gestalt", in *Psychologische Forschung*, 4, pp. 301-350, 1923.

[35] S. M. Williams, Perceptual Principles in Sound Grouping, in: G. Kramer (Ed.) *Auditory Display: Sonification, Audification, and Auditory Interfaces*, Reading, 1994, pp. 95–125

[36] M. Prensky, "Digital Natives, Digital Immigrants", in *On the Horizon*, MCB Univ. Press, vol. 9 no. 5, October 2001.

REVISITING PULSE RATE, FREQUENCY AND PERCEIVED URGENCY: HAVE RELATIONSHIPS CHANGED AND WHY?

Christian Gonzalez, Bridget A. Lewis, and Carryl L. Baldwin

George Mason University
4440 University Drive
Fairfax, VA
cgonza12@gmu.edu

ABSTRACT

Research in psychophysics and auditory warnings during the early 1990's created much of the theoretical groundwork for auditory alert design today. The main goal of this series of experiments was to reevaluate key auditory parameters (pulse rate and fundamental frequency) that have been shown to exhibit psychophysical relationships with perceived urgency in an updated context. Our results suggest that the relationship between pulse rate and perceived urgency may have weakened since the early 1990's, but the relationship between frequency and perceived urgency remains relatively stable. However, the relationship between pulse rate and perceived urgency was more reliable across multiple study manipulations relative to the relationship between frequency and perceived urgency. Based on its robustness across variable acoustic contexts, auditory alert designers wishing to convey a range of urgency levels may be more successful utilizing pulse rate rather than frequency.

1. INTRODUCTION

Sounds can capture people's attention no matter where they are looking. This makes the auditory modality well-suited for signaling events of varying criticality during visually demanding tasks like driving. The auditory environment within consumer and commercial vehicles is quickly becoming more heavily loaded with safety, communication and navigation technologies. This ongoing increase of in-vehicle technology requires manufacturers to develop auditory alerts that convey varying levels of urgency. Because sound is used to convey many different meanings, it becomes imperative to consider how various auditory parameters may impact perceptions of urgency and how appropriately matched these parameters are to the hazard levels they connote.

In 1993, Hellier, Edworthy and Dennis [1] demonstrated that Stevens' Power Law exponents [2] could be used to quantify the relationship between changes in auditory parameters and changes in perceived urgency. In their seminal paper, the authors had participants produce line ratings to represent the perceived urgency of sounds that varied systematically in different auditory parameters. These ratings were then used to create psychophysical functions (summarized by a Stevens' Power Law exponent) describing the relationship between each parameter and perceived urgency.

The power law exponents they identified successfully predicted the perceived urgency ratings of a new set of stimuli. Their results demonstrated that a) it is possible to assess the relationship between different auditory parameters and perceived urgency using psychophysical methods, and b) some parameters have a stronger relationship with urgency than others.

The impact of their work on auditory warning design has been substantial. This paper, as well as several other related papers [3–5], has served as the basis for the urgency mapping literature. This work has focused on systematically manipulating auditory parameters in a context neutral format. In a limited number of studies, urgency mapping has been examined within the context of driving [6–8].

However, in the nearly 20 years since the publication of Hellier et al.'s [1] original article, the prevalence and diversity of auditory alerts has increased dramatically. Given the increasingly complex soundscape, perceptions of various auditory parameters may have changed or may be influenced by concurrent changes in other dimensions. For example, increases in frequency may not seem as urgent if pulse rates are changing at the same time. We sought to examine these issues in the current investigation.

Stevens [9] has demonstrated that psychophysical judgments are relative to the set of stimuli being presented. As in-vehicle alerts can vary greatly in the information they represent, sets of alerts within vehicles will need to be heterogeneous [10], [11] in order for drivers to discriminate between the different intended meanings. Allowing participants to rate all levels of each parameter within the same experiment may help us better understand how a heterogeneous alert environment impacts ratings of perceived urgency.

1.1 The present studies

The primary goal of the present investigation was to replicate the basic psychophysical relationships between key auditory parameters (namely pulse rate and frequency) observed by

Hellier et al. [1]. Secondly, we wished to examine the impact of auditory context (presenting several different parameter changes within a single experiment) on perceived urgency by utilizing a within-subjects design. And finally, we sought to examine the impact of presenting the sounds within a driving context. Note that we made no attempt to simulate an actual driving context, but rather merely asked participants to consider how urgent the sounds presented would seem if heard while driving.

Psychophysical relationships between stimuli and subjective ratings have been shown to be consistent across cultures [12], [13], participants [14] and samples, [15–17] (see [18] for review) thus allowing for comparison across the four experiments presented here. The following studies systematically manipulated pulse rate, fundamental frequency and intensity because all three have been shown to relate to changes in perceived urgency [1], [3], [10]. We examined a number of methodological changes that impact the coherence between our results and those observed previously by Hellier et al. [1]. In general, we expected to find that increases in pulse rate and frequency would lead to higher ratings of perceived urgency. Based on Hellier et al.'s [1] results, we expected pulse rate to exhibit the strongest relationship with perceived urgency. We hypothesized that rating multiple types of stimuli at once could potentially cause participants to calibrate their ratings of perceived urgency based on a single, highly urgent sounding parameter.

2. METHOD (EXPERIMENT1)

2.1 Participants

Twenty-six graduate and undergraduate George Mason University students aged 18 to 25 (mean = 20.08; 12 female) voluntarily participated for class credit. All participants indicated they had normal hearing. A unique sample of participants was used for each experiment.

2.2 Design

A within-subjects designed was utilized. Each participant experienced and gave subjective ratings for all magnitude levels of all alerts. Each alert was presented three times within the experiment and all alerts were completely randomized. The average rating of each alert was used for analysis to mitigate any order effects.

2.3 Materials

2.3.1. Equipment

Alerts were presented in a sound attenuated laboratory on an Optiplex 745 Dell PC with a SoundMAX Integrated Digital HD Audio Driver Analog Device sound card. All alerts were presented through a pair of Sennheiser HD-280 stereo headphones. There was no evidence of intensity disparity between the left and right channel.

A MATLAB based program was written to present alerts as well as collect subjective ratings of urgency, annoyance and acceptability using a digital slider. The range of the slider included values between 0-100 and allowed participants to see their current rating. The program also allowed participants to adjust each rating until they felt it accurately reflected their perceptions before submitting.

2.3.2. Stimuli

A total of 21 stimuli were created, seven for each of the three auditory parameters that were investigated: fundamental frequency, intensity and pulse rate. Experiments 1 and 2 used stimuli that varied in all three parameters (21 total), but Experiments 3 and 4 used only stimuli that varied in pulse rate and fundamental frequency (14 total). Frequency and pulse rate alerts were created following the specifications of Hellier et al. [1], whereby varying durations of silence separated several standard "basic" pulses. The basic pulse used, based on the pulse-burst principles described by Patterson [19], was a 200 millisecond (ms) sine wave (20 ms on/offset) with a fundamental frequency (F_0) of 300 Hz and 15 harmonic components. Each alert was then made up of parametric variations of the basic pulse and varying durations of silence. Only one alert parameter was manipulated at a time while all other parameters were held constant to the basic pulse as described above. Unless intensity was being specifically manipulated, the basic pulse was presented at 75 dBA. This methodology ensured that our stimuli matched those used by Hellier et al. [1] exactly.

Table 1 provides a description of the stimuli used in the four experiments. The bolding within the columns indicates specifically what parameter changed in each stimulus. The seven fundamental frequency alerts consisted of six basic pulses of the same F_0 played in succession. There was no silence between the pulses and each alert had a total duration of 1200 ms. The 20 ms on/offset allowed the pulses to be discernible without the need for silence between pulses.

The seven intensity alerts varied in a similar fashion. Each alert consisted of six basic pulses with an F_0 of 300 Hz played in succession. Total duration for each alert was 1200 ms. Again, the on/offset allowed the pulses to be discerned without silence between each pulse. Using a Brüel & Kjær Sound Level Meter, we verified the intensity of each stimulus. Decibel measurements were taken from the individual pulses rather than the entire alert to avoid including the decreasing intensity of the onset and offset in our measurement.

The seven pulse rate alerts consisted of between four and twelve basic pulses (F_0 = 300 Hz) the inter-pulse interval (IPI) - or silence between pulses - varied from 475 to 9 ms. Pulse-to-pulse duration is defined as the duration from the start of one pulse to the start of the next pulse (pulse duration + IPI). The total time of each pulse rate alert approached, but did not exceed 2500 ms so each alert varied slightly in total duration. Pulse rate was derived via a formula based on one previously used by Hellier et al. [1]:

$$2500\text{ms/pulse-to-pulse time} \quad (1)$$

2500 ms represents the total approximate duration of each stimulus. 2500 ms was used as the total duration to standardize the rates for all pulse rate stimuli although the total true durations varied slightly. For example, a stimulus with a pulse rate of 3.69 would consist of four basic pulses of 200 ms each

	Level	# of Pulses	FFq (Hz)	IPI (ms)	Length (ms)	Intensity (dBA)
Pulse Rate	1	4	300	**475**	2225	75
	2	5	300	**302**	2210	75
	3	6	300	**238**	2389	75
	4	8	300	**118**	2430	75
	5	9	300	**60**	2280	75
	6	10	300	**50**	2450	75
	7	12	300	**9**	2499	75
Fundamental Frequency	1	6	**210**	0	1200	75
	2	6	**250**	0	1200	75
	3	6	**260**	0	1200	75
	4	6	**320**	0	1200	75
	5	6	**440**	0	1200	75
	6	6	**500**	0	1200	75
	7	6	**680**	0	1200	75
Intensity	1	6	300	0	1200	**66**
	2	6	300	0	1200	**69**
	3	6	300	0	1200	**72**
	4	6	300	0	1200	**75**
	5	6	300	0	1200	**78**
	6	6	300	0	1200	**81**
	7	6	300	0	1200	**84**

Table 1: Stimuli for Experiments 1 - 4. Experiments 1 and 2 used all 21 stimuli. Experiments 3 and 4 used only pulse rate and frequency stimuli.

separated by 475 ms of silence. Because following the last pulse was simply 275 ms of silence the total true duration of this alert is 2225 ms rather than 2500 ms.

2.4 Procedure

After completing an informed consent form, participants were told they would be presented with a variety of auditory alerts, which they would then rate on urgency, annoyance and acceptability. They were asked to imagine these alerts were presented in a driving context, but we did not specify in what capacity (e.g. collision warning, navigation, or communication etc.). We operationally defined acceptability as "How likely you would be to purchase a vehicle with this type of alert."

Participants then completed a brief practice with non-experimental auditory alerts to familiarize themselves with the rating slider. During the actual experiment, participants received a fixation cross on a black screen for 500 ms, then the auditory alert and a black screen, then three separate rating screens for urgency, annoyance and acceptability. The rating screen order was consistent throughout the experiment. The experiment took approximately 30 minutes to complete.

3. RESULTS AND DISCUSSION (EXPERIMENT 1)

Though annoyance and acceptability results were analyzed in Experiments 1 - 4, they will not be discussed here. The goal of this paper was to re-examine Hellier et al.'s [1] findings, which pertained only to ratings of perceived urgency. See [18 *under review*] for a closer examination of urgency and annoyance ratings.

Results were analyzed according to Hellier et al.'s [1] specifications. Exponents were calculated for each parameter according Stevens' [2] methodology. All raw values were log transformed and their geometric means were taken. All parameter values were also log transformed. This allowed us to create a log-log plot of perceived urgency ratings as a function of changes in each parameter. The slope of the best-fit line plotted through these points is the exponent used in Stevens' Power Law:

$$P = kS^n \quad (2)$$

Where P is the perceived urgency rating of the physical stimulus (S), k is a constant and n is the exponent found using empirical data. Smaller exponents are related to smaller changes in perceived urgency as the stimulus changes whereas larger exponents (generally greater than 1) are related to larger changes in perceived urgency relative to stimulus changes. Similar to Hellier et al.'s [1] findings, this experiment also found a large portion of the variance could be accounted for by the slope of a best fit line (see Table 2). This supports Hellier et

al.'s [1] assertion that it is possible to systematically quantify changes in perceived urgency with relation to changes in parameter level. The percent variance explained in Table 2 is in reference to the variance explained among the seven log-transformed mean values of each parameter level, not the variance explained among all participants' ratings.

As illustrated in Table 2, we found that intensity produced the largest exponent (n = 3.8) while pulse rate (n = .52) and frequency (n = .54) produced similar, much smaller exponents. Our pulse rate exponent was nearly 60% smaller than Hellier et al.'s [1] (see Table 6) indicating a weaker relationship with urgency than expected. However, our frequency exponent was slightly larger than Hellier et al.'s [1] findings.

Level	Pulse Rate	Fundamental Frequency	Intensity
	Mean rating value (0-100) and standard deviation		
1	46.61 (21.5)	46.90 (19.4)	40.70 (22.5)
2	55.59 (17.6)	53.54 (18.5)	54.33 (20.7)
3	64.90 (16.3)	55.06 (16.8)	62.82 (19.6)
4	71.97 (15.4)	70.70 (14.8)	69.18 (14.6)
5	75.11 (13.6)	69.19 (18.8)	74.23 (10.81)
6	72.75 (14.3)	74.13 (15.6)	78.90 (10.7)
7	78.61 (12.4)	73.52 (19.3)	84.55 (9.4)
Exponent	0.51	0.54	3.8
% Variance accounted for	0.91	0.75	0.92

Table 2: Experiment 1 - Effects of Three Auditory Parameters on Perceived Urgency.

4. METHOD (EXPERIMENT 2)

4.1 Introduction

Experiment 2 was very similar to Experiment 1 with the exception that we provided an additional visual cue to better connote a driving context for participant ratings.

4.2 Participants

Thirty-one graduate and undergraduate George Mason University students aged 18 to 25 (mean = 19.5; 22 female) voluntarily participated for class credit.

4.3 Procedure

Experiment 2 followed the exact specifications of Experiment 1, except that instead of a black screen with a fixation cross, participants saw a generic car dashboard on the screen.

5. RESULTS (EXPERIMENT 2)

Results were examined using the exact procedure described for Experiment 1. Again, we found intensity alerts produced the largest exponent (n = 2.6) by far (see Table 3), though 30% smaller than Experiment 1. Pulse rate alerts produced a similar exponent as Experiment 1 (n = .47) while frequency alerts produced a much smaller exponent (n = .29). These findings still differ greatly from Hellier et al.'s [1] results (Table 6) and may suggest that the relationship between frequency and perceived urgency may be more sensitive to even small changes in context than pulse rate. The fact that intensity produced such a large exponent could be indicative participants calibrating their ratings of perceived urgency for pulse rate and frequency alerts. Pulse rate and frequency could have been perceived as less urgent in the context of another seemingly much more urgent sounding alert (intensity alerts). While the relationship between intensity and perceived urgency is seemingly quite strong, if we wish to maintain guidelines established by Patterson [17] [i.e. warnings should be presented at least 15 decibels (dB) above ambient background noise], intensity would likely not be a feasible parameter to manipulate in a noisy vehicle. For Experiment 3, we eliminated intensity from our manipulations and examined the impact of experiencing changes in pulse rate followed by frequency on ratings of perceived urgency.

Level	Pulse Rate	Fundamental Frequency	Intensity
	Mean rating value (0-100) and standard deviation		
1	52.31 (27.9)	60.93 (23.8)	53.22 (26.1)
2	57.72 (26.4)	64.10 (22.8)	60.81 (23.6)
3	64.61 (22)	65.19 (22.5)	66.28 (20.7)
4	72.47 (21.8)	71.31 (18.1)	70.76 (17.7)
5	74.54 (21.4)	70.47 (19.9)	74.49 (16.9)
6	74.39 (18.6)	73.63 (17.7)	80.57 (11.3)
7	77.32 (18.3)	75.11 (18.4)	84.31 (12.8)
Exponent	0.47	0.28	2.6
% Variance accounted for	0.94	0.87	0.97

Table 3: Experiment 2 - Effects of Three Auditory Parameters on Perceived Urgency.

6. METHOD (EXPERIMENT 3)

6.1 Introduction

Based on the results of Experiment 2, we split Experiment 3 into blocks. Block 1 always consisted of only pulse rate alerts and Block 2 always consisted of only frequency alerts. Block order was not manipulated and participants were not made aware of the block changes. This was done to mitigate any rating calibration effects discussed in Experiment 2. Also, this more closely mirrors Hellier et al. [1] where they ran four individual smaller experiments to collect ratings of urgency.

6.2 Participants

Thirty graduate and undergraduate George Mason University students aged 18 to 29 (mean = 20.52; 6 female) voluntarily participated for class credit. All participants indicated they had normal hearing.

6.3 Equipment

Due to our inability to closely reproduce Hellier et al.'s [1] findings, we decided to change our rating scale to more accurately mimic the paper and pencil methodology they used. Instead of a digital slider, participants provided ratings via an on-screen line draw. Participants could draw a straight, horizontal line anywhere on the rating screen using the mouse. The maximum possible length was the equivalent of 394 millimeters (the maximum possible line length in Hellier et al.'s [1] study). Participants were not given feedback on the magnitude of their ratings. The length of the line was recorded in pixels then converted to mm to more accurately reflect the data and scale used in Hellier et al. [1].

This change in scale coupled with a change in parameters investigated is not an ideal manipulation. However, [9] has demonstrated that, in general, relationships between stimuli and ratings are independent of the rating scale. Thus, we combined the manipulations in order to constrain the number of experiments in this series.

6.4 Procedure

Other than the changes noted above, the procedures were identical to Experiment 1 and 2. We verified that participants understood that the length of the line reflected the magnitude of their rating, such that longer lines meant alerts were more urgent, more annoying, and more acceptable and vice versa. The total experiment time was reduced to 20 minutes.

7. RESULTS AND DISCUSSION (EXPERIMENT 3)

The change in rating scale from Experiments 1 and 2 to Experiment 3 necessitated a slightly lengthier transformation in order to compare across studies. Ratings were converted from mm on the screen to a percentage of the total possible mm rating they could have given, thus allowing for comparison between mm ratings and slider ratings out of 100. These converted percentage values were then log transformed to derive the exponents. The methods used in Experiment 3 resulted in a much larger exponent relative to Experiments 1 and 2 for pulse rate alerts (n = .77). However, frequency alerts produced a non-significant exponent (n = .10). Once again we were unable to closely replicate Hellier et al.'s 1993 [1] findings. The dramatic change in the exponent for frequency seemed likely due to an order effect. Participants may have calibrated their ratings of frequency relative to the block in which pulse rate was manipulated. The increase in the observed exponent for pulse rate in Experiment 3 supports our previous suspicion that exposure to sounds varying in intensity (Experiments 1 and 2) resulted in some rating compression. The methodology in Experiment 3 mirrors Hellier et al. [1] in stimuli, rating method and (pseudo) between-subjects design more so than of the previous studies. Although our pulse rate exponent is closer to their original findings, it is still over 40% smaller. This may indicate that even under nearly identical conditions, the relationship between pulse rate and perceived urgency has changed over the last 20 years.

8. METHOD (EXPERIMENT 4)

8.1 Introduction

In order to examine the potential order effects from Experiment 3, we ran Experiment 4 with a reversed block order where Block 1 consisted of only changes in frequency and Block 2 consisted of only changes in pulse rate.

Level	Pulse Rate	Fundamental Frequency
	Mean rating value (%) and standard deviation	
1	29.08 (16.7)	31.50 (20.1)
2	34.09 (19.3)	34.89 (22.3)
3	38.58 (20.2)	31.54 (22.7)
4	47.66 (22.3)	36.12 (20.5)
5	52.37 (23.5)	37.93 (24.5)
6	56.31 (24.4)	37.91 (30.2)
7	64.20 (26.1)	43.05 (32.4)
Exponent	0.47	0.28
% Variance accounted for	0.94	0.87

Table 4: Experiment 3 - Effects of Three Auditory Parameters on Perceived Urgency.

8.2 Participants

Ten graduate and undergraduate George Mason University students aged 18 to 22 (mean = 19.12; 13 female) voluntarily participated for class credit. All participants indicated they had normal hearing.

9. RESULTS AND DISCUSSION (EXPERIMENT 4)

Results were analyzed using the same procedures employed in Experiment 3. We observed a smaller exponent for manipulations of pulse rate (n = .50) and a much larger exponent for frequency (n = .38) indicating there may have been some block order effects on ratings of perceived urgency. However, our exponent for frequency matched Hellier et al.'s [1] findings almost exactly. This suggests that changes in frequency may still have a similar relationship with perceived urgency though only under specific and homogenous conditions.

10. GENERAL RESULTS

Table 6 summarizes exponent values, effect sizes and 95% Confidence Intervals (CIs) for pulse rate and frequency across

all four studies and Hellier et al. [1]. In order to investigate differences across studies we utilized 95% CIs of the exponents derived from the log-log regression plots. Though 95% CIs were not reported in their original article, Hellier et al. [1] did provide raw data from their experiments. This allowed us to analyze their data and identify the CIs for pulse rate and frequency exponents. We converted their raw millimeter values to percentages and then log transformed them, similar to Experiments 3 and 4, allowing for comparison on a 0-100 scale across all experiments.

Level	Pulse Rate	Fundamental Frequency
	Mean rating value (%) and standard deviation	
1	34.46 (18.5)	32.27 (16.6)
2	32.86 (8.2)	35.93 (18.6)
3	46.19 (20.9)	38.81 (22.2)
4	53.29 (22.5)	37.53 (22.1)
5	51.11 (17.1)	45.50 (25)
6	50.22 (23.43)	48.29 (24.4)
7	58.23 (20.7)	53.19 (29)
Exponent	0.50	0.38
% Variance accounted for	0.87	0.88

Table 5: Experiment 4 - Effects of Two Auditory Parameters on Perceived Urgency.

Because of the variation in samples and methodologies we encourage caution when interpreting the CIs across the four experiments and Hellier et al. [1]. Finding statistically significant differences was not the ultimate goal for this series rather exploration of previous relationships. Table 6 also includes R^2 values reported in Hellier et al. [1] for comparison purposes. (R^2 values are equivalent to the percent of variance accounted for).

10.1 Pulse rate across experiments

We found the only experiment that did not fall within the 95% CI of another was Experiment 3. This experiment produced the largest exponent for pulse rate (n = .77) and falls outside of the CI of Experiment 2. Experiment 3 also approached the upper limits of Experiment 1 and 4's CIs. However, overall, exponents from the four experiments remained quite similar. In comparison none of the exponents from the four experiments fell within the 95% CI of Hellier et al.'s [1] exponent value, indicating that the relationship between pulse rate and perceived urgency may have weakened over time.

Figure 1 shows a log-log plot of pulse rate on the x-axis and average perceived urgency rating on the y-axis. The mean rating values for each level of pulse rate are shown with a line of best fit for each experiment and Hellier et al. [1]. The exponents reported in Table 6 reflect the slope of each line. This figure highlights the similarity in slopes across all four experiments. The y-intercepts for Experiments 3 and 4 differ from Experiments 1 and 2 because of the difference in scales between the two sets of studies. However, a general slope characteristic is maintained by the four experiments illustrating the results reported in Table 6. In comparison, the slope of the line of best fit from Hellier et al. [1] appears much steeper demonstrating the large difference in exponents.

10.2 Frequency across studies

We found a much larger range of exponents for frequency across all four experiments. We also found that no single experiment fell outside the 95% CI of any other. However, when looking at the exponent values we see a much larger spread for frequency than pulse rate. We also see consistently

Parameter	Exp.	n	R^2	*p*	95% Lower	95% Higher
Pulse Rate	1	.51	.91	0	.33	.71
	2	.47	.94	0	.33	.62
	3	.77	.99	0	.70	.85
	4	.50	.87	0	.28	.73
	Hellier et al. [1]	1.35	.98	0	1.16	1.54
Frequency	1	.54	.75	.01	0.18	.90
	2	.28	.87	0	0.16	.42
	3	.10	.15	.40	NS	NS
	4	.38	.88	0	0.23	.55
	Hellier et al. [1]	.38	.93	0	.29	.54

Table 6:Summary results Experiment 1 – 4 and Hellier et al. [1]

smaller R^2 values for frequency across experiments. Only three of the four experiments found a relationship between changes in frequency and perceived urgency. However, all four experiments fall within the 95% CI of Hellier et al.'s [1] exponent value. In addition to producing the same exponent, Experiment 4 also produced upper and lower CI bounds similar to Hellier et al. [1]. This may indicate that the relationship between frequency and perceived urgency has changed less than pulse rate over time. Figure 2 shows a log-log plot similar to Figure 1, but with fundamental frequency on the x-axis. Figure 2 illustrates the variation in slopes across all 4 experiments as well as the greater variance of mean data points compared to Figure 1. The similarity in slopes of Experiment 4 and Hellier et al. [1] is also evident though the y-intercept values differ due to scale differences.

11. GENERAL DISCUSSION

Our findings indicate that across various procedural and methodological manipulations and within homogenous and heterogeneous alert sets, pulse rate exhibits a relatively robust relationship with perceived urgency. However, the magnitude of this relationship may have weakened since Hellier et al.'s 1993 [1] experiment 20 years ago. Though it is difficult to systematically evaluate the role that increased technology and

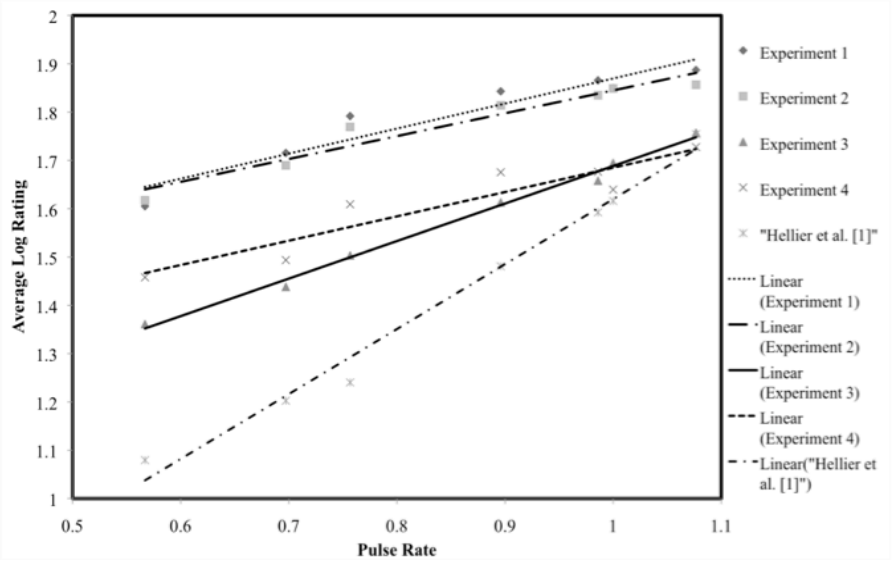


Figure 1: Log-log plot of pulse rate and ratings of perceived urgency across four experiments.

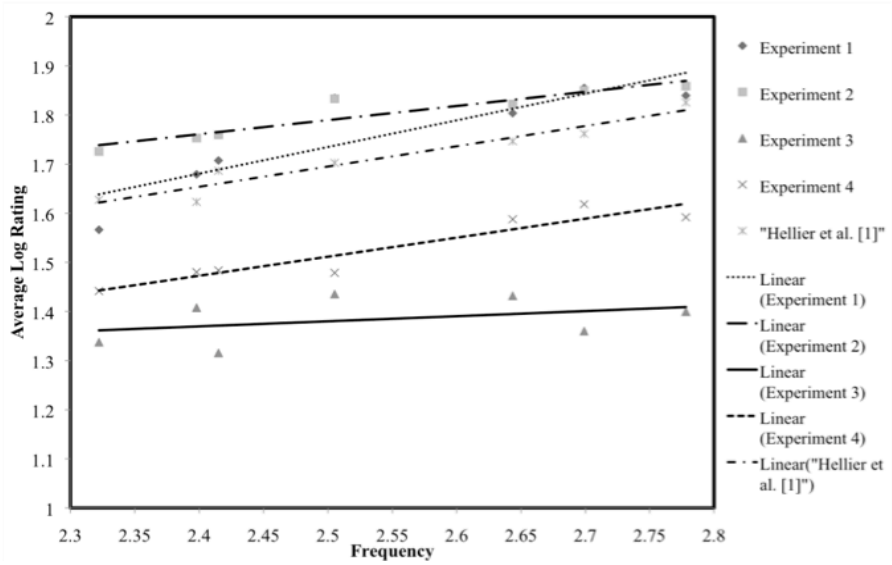


Figure 2: Log-log plot of frequency and ratings of perceived urgency across four experiments.

sound exposure plays in this weakening, it seems plausible that sensitivity to changing pulse rates has decreased with a general increase in exposure to technology. The stability of pulse rate over multiple studies is in agreement with Patterson's [19] finding that temporal patterns are the critical structural difference when distinguishing between sounds. Furthermore, the robustness of pulse rate across manipulations may also be explained by the ability of the auditory system to distinguish minute changes in timing in concert with other highly variable parameters, as exhibited by the role of temporal characteristics in perception of phonemic changes resulting from coarticulation and changes in Voice Onset Time on the millisecond level [21], [22].

The relationship between frequency and perceived urgency seems to be conditional on the presence of other alerts against which it may be compared. Furthermore, rating order also seems to have a large impact on the relationship between frequency and perceived urgency. Hellier et al. [1] suggested a similar unreliability in the frequency exponent they reported claiming the metathetic nature of frequency [23] as a potential explanation. Previous research [24] has also shown the ability to retain pitch decays over time and is subject to interference from other tones [25]. This may make frequency less than ideal for conveying multiple levels of urgency especially in a heterogeneous environment. Surprisingly, while the relationship of pulse rate and perceived urgency seems to have weakened over time, frequency, under homogenous conditions (Experiment 4), was the only parameter that produced a similar power law exponent to those reported by Hellier et al. [1].

Though different samples of participants were used for

each experiment, both pulse rate and frequency ratings were subject to the same potential influence of individual differences. As [26] has shown, individual variation in power law exponents is indeed random and pooling subject data results in reliable exponents over time and across samples. Barring that, there is still a chance that discrepancy in methods represents some of the variation in power law exponents found between studies. However, across manipulations and ostensibly different samples, pulse rate maintained a relatively stable exponent clearly different from Hellier et al.'s [1] results.

The two main findings applicable to auditory alert designers are: 1) Alerts within a heterogeneous set (similar to what may be found in vehicles or other complex auditory environments) may have different relationships with perceived urgency than those same alerts in a homogenous set. 2) When it is critical to convey a specific level of urgency aurally, pulse rate may be the most reliable and robust parameter to manipulate. However, due to the apparent weakening of the magnitude of the relationship between pulse rate and perceived urgency, increased levels of pulse rate may be needed to convey the same level of urgency achieved 20 years ago.

Together, the present series of experiments examined some of the many methodological factors that may impact the relationship between auditory parameters and perceptions of urgency. The relationship between pulse rate and perceived urgency appears to have changed over time, but it remains well-suited for use in the design of effective in-vehicle alerts and alarms. In the future, we plan to extend this work into higher fidelity simulations where we can evaluate the impact of more realistic driving contexts on ratings of perceived urgency.

12. ACKNOWLEDGEMENTS

The authors would like to thank Stephanie M. Pratt and Daniel M. Roberts for making this work possible through their assistance in coding and developing the various study iterations.

13. REFERENCES

[1] E. Hellier, J. Edworthy, and I. Dennis, "Improving auditory warning design: Quantifying and predicting the effects of different warning parameters on perceived urgency," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 35, no. 4, pp. 693–706, 1993.

[2] S. S. Stevens, "On the psychophysical law.," *Psychological review*, vol. 64, no. 3, p. 153, 1957.

[3] E. Hellier and J. Edworthy, "On using psychophysical techniques to achieve urgency mapping in auditory warnings," *Applied Ergonomics*, vol. 30, no. 2, pp. 167–171, 1999.

[4] J. Edworthy, S. Loxley, and I. Dennis, "Improving auditory warning design: Relationship between warning sound parameters and perceived urgency," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 33, no. 2, pp. 205–231, 1991.

[5] E. Hellier, J. Edworthy, and I. Dennis, "A comparison of different techniques for scaling perceived urgency," *Ergonomics*, vol. 38, no. 4, pp. 659–670, 1995.

[6] C. L. Baldwin, "Verbal collision avoidance messages during simulated driving: perceived urgency, alerting effectiveness and annoyance," *Ergonomics*, vol. 54, no. 4, pp. 328–337, Apr. 2011.

[7] C. L. Baldwin and J. F. May, "Loudness interacts with semantics in auditory warnings to impact rear-end collisions," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 14, no. 1, pp. 36–42, Jan. 2011.

[8] D. C. Marshall, J. D. Lee, and P. A. Austria, "Alerts for in-vehicle information systems: Annoyance, urgency, and appropriateness," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 49, no. 1, pp. 145–157, 2007.

[9] S. S. Stevens, "On the operation known as judgment," *American Scientist*, vol. 54, no. 4, pp. 385–401, 1966.

[10] J. Edworthy, E. Hellier, K. Titchener, A. Naweed, and R. Roels, "Heterogeneity in auditory alarm sets makes them easier to learn," *International Journal of Industrial Ergonomics*, vol. 41, no. 2, pp. 136–146, Mar. 2011.

[11] M. A. Nees and B. N. Walker, "Auditory Displays for In-Vehicle Technologies," *Reviews of Human Factors and Ergonomics*, vol. 7, no. 1, pp. 58–99, Aug. 2011.

[12] E. C. Haas and J. Edworthy, "Designing urgency into auditory warnings using pitch, speed and loudness," *Computing & Control Engineering Journal*, vol. 7, no. 4, pp. 193–198, 1996.

[13] C. Ellen and J. G. Casali, "Perceived urgency of and response time to multi-tone and frequency-modulated warning signals in broadband noise," *Ergonomics*, vol. 38, no. 11, pp. 2313–2326, 1995.

[14] Chillery, J. A. and Collister, J. B., "Possible confusions amongst a set of auditory warning signals developed for helicopters," Farnborough: Royal Aircraft Establishment, Technical Memo FS(F) 655, 1986.

[15] B. N. Walker, "Consistency of magnitude estimations with conceptual data dimensions used for sonification," *Applied Cognitive Psychology*, vol. 21, no. 5, pp. 579–599, Jul. 2007.

[16] J. Edworthy and S. Newstead, "On the Stability of the Arousal Strength of Warning Signal Words," 1999.

[17] M. Teghtsoonian, "Children's scales of length and loudness: A developmental application of cross-modal matching," *Journal of Experimental Child Psychology*, vol. 30, no. 2, pp. 290–307, Oct. 1980.

[18] E. Hellier and J. Edworthy, "Subjective rating scales: scientific measures of perceived urgency?," *Ergonomics*, vol. 45, no. 14, pp. 1011–1014, 2002.

[19] R. D. Patterson, *Guidelines for auditory warning systems on civil aircraft*. Civil Aviation Authority, 1982.

[20] Gonzalez, C., Lewis, B. A., Pratt, S. M., Roberts, D. M., and Baldwin, C. L., "Perceived urgency and annoyance of auditory alerts in a driving context. (under review)," in *Proceedings from HFES '12*, Boston, MA, 2012.

[21] L. Lisker and A. S. Abramson, "A cross-language study of voicing in initial stops: acoustical measurements," *Word*, vol. 20, no. 3, pp. 384–422, 1964.

[22] W. Strange, R. R. Verbrugge, D. P. Shankweiler, and T. R. Edman, "Consonant environment specifies vowel identity," *The Journal of the Acoustical Society of America*, vol. 60, no. 1, pp. 213–224, 1976.

[23] S. S. Stevens and E. H. Galanter, "Ratio scales and category scales for a dozen perceptual continua.," *Journal of Experimental Psychology*, vol. 54, no. 6, p. 377, 1957.

[24] D. Deutsch, "Delayed pitch comparisons and the principle of proximity," *Perception & Psychophysics*, vol. 23, no. 3, pp. 227 – 230, 1978.

[25] D. Deutsch, "The processing of pitch combinations," *The psychology of music*, pp. 271–316, 1982.

[26] M. Teghtsoonian and R. Teghtsoonian, "How repeatable are Stevens's power law exponents for individual subjects?," *Attention, Perception, & Psychophysics*, vol. 10, no. 3, pp. 147–149, 1971.

EVERYDAY LISTENING TO AUDITORY DISPLAYS: LESSONS FROM ACOUSTIC ECOLOGY

Milena Droumeva

Faculty of Education,
Simon Fraser University
mvdroume@sfu.ca

Iain McGregor

School of Computing,
Edinburgh Napier University
i.mcgregor@napier.ac.uk

ABSTRACT

In order to design auditory displays that function well within the cultural, informational and acoustic ecology of everyday situations designers as well as researchers in psychoacoustics need to continue to gain a better understanding of how listeners hear and make sense of information in more ecological settings and outside the lab! In this paper the authors present a preliminary study that builds on past work and theoretical ideas from acoustic ecology, exploring the practice of *everyday listening* in settings containing auditory displays. This pilot study involves 10 participants who are asked to listen to two separate soundscapes and describe in three tasks, both verbally and in writing, what they hear and how they make sense of these aural environments. The results suggest directions for understanding everyday listening form a holistic perspective in order to inform both the design of auditory displays, and the development of other research tools and instruments for measuring auditory perception ecologically. The bigger study which involves 100 participants has been completed and is expected to be published shortly as a journal article.

1. INTRODUCTION

As auditory displays become increasingly integrated within everyday products, services and environments, both designers of auditory displays and researchers of auditory perception have to continue to find better ways of understanding how these new ecologies of listening and sonic messages function together. Laboratory experiments with simple tones, while useful in establishing baseline psychoacoustic guidelines, become more and more insufficient in addressing listening as an everyday practice given the widening gap between psychoacoustic research and 'everyday' settings. As interdisciplinary research begins to become the norm rather than the exception in exploring and researching complex phenomena, the authors hereby attempt to infuse and mobilize several fields of study towards the investigation of everyday listening. In particular, we suggest that acoustic ecology offers some useful frameworks for understanding how soundscapes function ecologically and how listeners approach the reception and interpretation of sonic messages within their larger acoustic environment, including its socio-cultural context, informational and semiotic ecologies. The study we present here offers a preliminary attempt to identify salient themes, approaches and ways of mapping complex, everyday soundscapes that contain auditory displays, through both linguistic, reflective, and graphic notation systems. For this pilot study, we begin with linguistic and narrative structures and in analyzing them, help identify, categorize and develop ways of representing the various sonic, spatial and temporal elements of a given (electro)-acoustic ecology.

2. SOUNDSCAPE MAPPING – PAST RESEARCH

The need for developing multi-lateral tools for soundscape mapping in research that aims at understanding how auditory displays fit in and function within complex "everyday" environments has already been documented [1, 2]. However, initiatives to understand listening, outside of its purely perceptual and psychological characteristics, are few to find. Fewer still are examples of studies where soundscape mapping is connected explicitly with notions from acoustic ecology. We believe it is crucial, particularly in our increasingly ambient intelligent multi-sensory environments that research should aim at exploring more *ecological* notions of listening and focus on how people attend to and make sense of their everyday soundscapes. Such studies would focus on two aspects of auditory display research – firstly, on improving the ecological validity of psychoacoustic research by infusing it with frameworks and approaches such as acoustic ecology (but potentially open to cultural and critical approaches as well); and secondly, by continuing to develop soundscape mapping/research methodologies and identifying salient perceptual characteristics for the reception of auditory displays in everyday contexts.

Soundscape mapping can take the form of various graphic notation systems for logging and representing both individual sounds and entire soundscapes. It exists as a tool in several areas of research, design and community practice: classifying the elements of a soundscape – a type of comprehensive auditory ontology – through either a functional/categorical or spatially-oriented framework; visualizing soundfield measurements and sonic characteristics such as magnitude, frequency spectrum, dynamics and temporality; and finally, representing a listener's perspective of a given soundscape. Classifying sonic elements is not new – important past works include Gaver's [3] classification of everyday sounds as well as Hellström's [4] mapping schema combining spatial and structural sonic components. Organizing soundscape classifications according to perceived sound quality, aesthetic or emotional content, spatial characteristics, interactive functionality and informational significance has resulted in a

number of soundscape ontologies that remain in schematic form. Notation systems that progress to graphical representation include Coleman, Macauley and Newell's [5] sound map tool designed for participatory workshops, similar design process instruments and most notably the tools, frameworks and classifications to come from the ethnographic work of R. M. Schafer [6] and the World Soundscape Project in the late 1960s/early 70s. Schafer's approach to soundscape mapping is most unique in the ecological framework within which sound is positioned as a subject of study and as a phenomenological experience. Schafer's classification of soundscape components into prominent or significant sonic characteristics that define communities reflects a view of soundscapes as profoundly listener-centered. In other words, the significance of each sound environment, each context in which a variety of sounds exist in an "acoustic ecology" is determined and shaped by the listeners who occupy that setting. This represents a shift in soundscape mapping frameworks from ones that focus largely on the informational and functional characteristics of sounds, to ones that focus on people's listening experiences in various degrees of complexity. Again, this is critical, we think, to understanding the context in which people experience auditory displays in everyday life, both in specific situations, as we all in terms of macro trends of listening attention, information retrieval, and other associative characteristics inside a perception-cognition-action loop. Such approaches, naturally, also have predecessors. In surveying the field of what Schiewe and Kornfield [8] refer to as audio cartography, they argue that the visualization of sound has been for the most part disregarded and limited in scope. They suggest that acoustic geography should incorporate both subjective and measured dimensions in spatial terms, and that descriptions and measurements ought to be combined from soundscape research [6], acoustics [9] and psychoacoustics [10]. Their system identifies the following elements to be incorporated in soundscape mapping: sonic balance, sound events and soundmarks; sound pressure levels, intensity, trajectory and frequency of sound; perceptual parameters such as loudness, pitch, timbre, rhythm, fluctuation and annoyance. Figure 1 shows an illustration of employing these elements in a graphed sound zone, and a dynamic listener profile alongside.

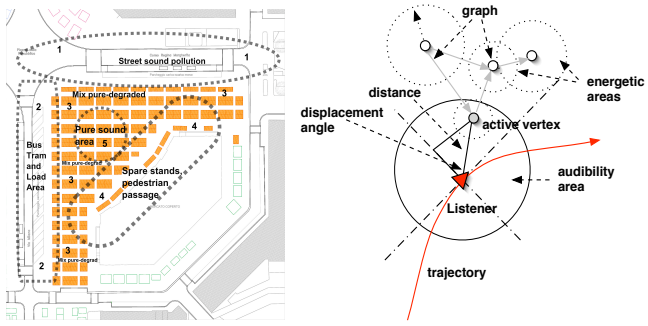


Figure 1: Map indicating sound zones and a listener profile including direction and trajectories of auditory horizon [11].

Within the category of soundscape mapping as a perceptual phenomenon, we distinguish techniques that only include listeners' experiences [7, 12, 13, 14, 15] from methods that combine both soundfield measurements and listeners' experiences [6, 8, 11, 16]. Most approaches rely predominantly on the identification and meaning of sound sources along with spatial, dynamics, temporal and spectral attributes. Most works are preliminary and often lack fully annotated examples, or simply do not provide a basis for their graphical, aesthetical and functionally representative choices for soundscape mapping. Thus one of the critical tasks scaffolding any attempt to develop a comprehensive system of soundscape mapping is a good classification system of both the soundscape's elements, as well as the relevant aspects of listening. Identifying what are important characteristics about the soundscape and about the listening experience in everyday contexts is precisely the gap that requires further exploration in order to inform both the fields of auditory display design and auditory perception research.

3. LESSONS FROM ACOUSTIC ECOLOGY

The main reason for harnessing acoustic ecology in auditory display research is of course to better understand the complexities of listening and to help develop more comprehensive tools for mapping soundscapes both in terms of how auditory displays fit in a given environment/context, and how people listen to and make sense of these augmented environments. Acoustic ecology is a field of study, research and international activism that was established through Schafer's [6] work with the World Soundscape Project (WSP). Concerns over rising urban noise levels and a commitment to preserving the participatory and communal nature of the acoustic environments are at the heart of acoustic ecology. That project – the result of several years' worth of ethnographic work mainly located around five villages in Western Europe – reveals, among other things, strong connections between the aural world, local culture and the functioning of everyday life. This is documented in numerous interviews with local residents about their soundscapes revealing a deep relationship between the aural environment and notions of place, time and self. In publications following the WSP, and with the help of the WSP team, Schafer developed a simple organizing ontology of the soundscape as containing at least three types of sounds – signals, soundmarks and keynote sounds [6]. While these sounds would be different for each 'acoustic community' (see Figure 3) depending on what sounds take on significance in the local soundscape, they would *function* in similar ways everywhere. Soundmarks in particular, termed after visual landmarks, are sounds that listeners associate strongly with their acoustic community – examples could be anything from factory steam whistles, to water streams, church bells and typical bird songs [6]. Acoustic communities are not static, however, as significant sounds become introduced in the soundscape, they change and shift in importance with time. It is the listeners and their awareness and acknowledgement of the emplacing, situational nature of sound that supplies the other ingredient of each acoustic ecology. As Truax [18] points out, extending Schafer's notions of the soundscape, the nature of acoustic communication positions the listener, the sound and the soundscape in a dynamic, two-way flow of interaction, communication and interdependence. Both our listening and our soundmaking, according to Truax, are functions of the context in which we listen and sound – not only culturally, but literally. Our ears pick up on relevant cues and properties of each (electro)-acoustic context in order to apply dimensional and associational judgments and sort of what sound events,

sound characteristics and informational aspects to tune into [18]. One simple example is the concept of acoustic *masking* – when we are in an environment, which is 'loud,' we have to respond by raising our voices in order to communicate, in essence adding to the noise. However, in a more granular aspect of that situation, one that Truax terms 'cocktail-party effect' our ears pick up on the voice of a familiar person even in the crowd and noisiness of a group event. In acoustic ecology, special importance is placed on the distinction between acoustic and electroacoustic environments. Marked by the possibility of artificial amplification, which necessarily shifts the sonic balance of natural environments, electroacoustic communication [18] entails cultural, social and economical dimensions in the way it acculturates listeners. Exposure to media soundscapes for over a century now has given rise, according to Truax, to a variety of specific listening positions that are attuned to and respond to the flow, construction and sonic parameters of media listening [18]. Yet there is a redeeming factor in the notion of an 'ecology' that resists a technological determinism that typically blames sonic imbalance on technological urban progress. In fact, Truax and Schafer would argue that even by virtue of being and acting in a soundscape, we affect it as both its listeners and its composers [6, 18].

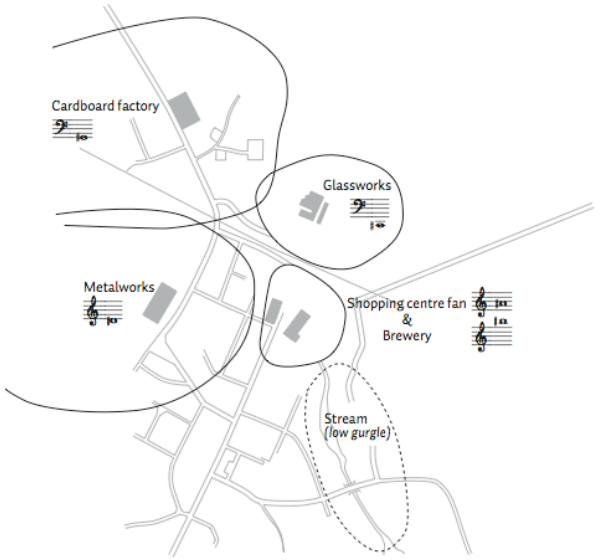


Figure 2: Diagram of the acoustic profiles of local soundmarks and keynotes from the village of Skriv, reprint from Acoustic Environments in Change/ Five Village Soundscapes [19].

The notions most important to our present project to come out of acoustic ecology involve three ideas: graphically representing – soundscape mapping – multiple listener accounts, that is to say, presenting a macro scale of soundfield information and listener data – see Figure 2 – while placing special importance on the layers of sound information, the sound profiles (audible scopes) of various elements and the way in which they constitute particular *electroacoustic communities*. The significance of this approach to auditory display research is, of course, the fact that, ecologically-speaking, there are not only multiple auditory displays in a given setting, but there are normally multiple listeners that researchers, as well as designers, rarely explore on a macro-level, thus obscuring the communal experience of hearing and interpreting auditory displays in the context of each electroacoustic community. More contemporary work at the intersection of acoustic ecology and cultural studies serves as proof that the potential of this field is yet to fully blossom [12]. Another useful notion to come out of the acoustic ecology field is a sensitivity to the temporal dimension of listening. While much of auditory perception research and auditory display design assumes that sounds are experienced fundamentally on a spatial plane in a single unit of time; and that soundscapes are place-bound [4, 13, 14] everyday listening is essentially temporal, event-related, intimately coupled with context, subjectivity and attention – which are purveyors of time as well. Space and time, therefore must both be accounted for in an ecological instrument for understanding everyday listening. For designers and researchers of auditory displays, it is not, perhaps, quite enough to understand how well listeners can spatially locate as well as functionally and informationally identify sound signals – it is also important to understand how listening shifts and how soundscapes themselves change, both ecologically and perceptually, over time. The graph in Figure 3 also comes from Five Villages, part of the WSP project [6], and reflects a graphic combination of sound level measurement with time coded, annotated sound events. Many more such hand-drawn graphics and maps can be found in the supplementary WSP materials library. Sound graphs, as well as sonic maps, are integral ways of representing a soundscape as a listening account, while being sensitive to both temporal and spatial dimensions of soundscapes and of listeners.

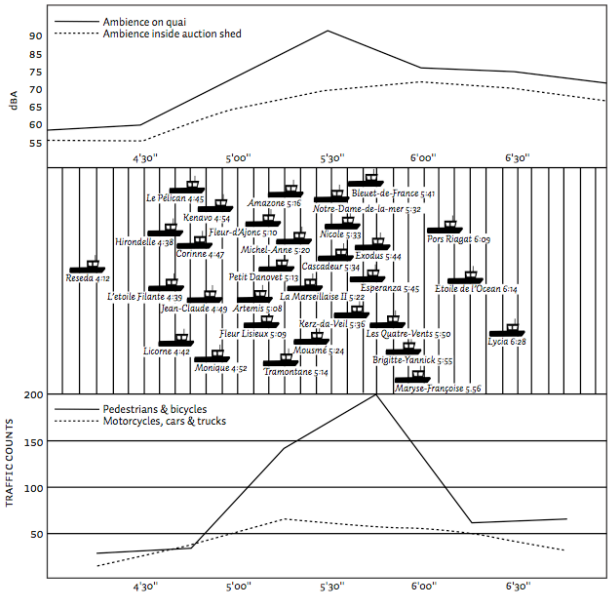


Figure 3: A temporal sound diagram reflecting the arrival of the fishing boats in Lesconil, France prior to the daily auction, as documented in Five Village Soundscapes - reprint from Acoustic Environments in Change, [19].

Finally, a method from Schafer's work that has been used in others' design work already, and which, admittedly, resembles

similar ethnographic approaches is the *earwitness account*. An interview elicited specifically with regard to a regular listener's intimate familiarity with the soundscape, in some reflective detail constitutes an earwitness account. While Schafer didn't explicitly acknowledge it, much of his background research relies on language, particularly literary accounts of historical soundscapes [6]. While language is limited in the sense that untrained listeners rarely possess a great vocabulary to describe their soundscape (Schafer imagines a long programme of ear-cleaning and re-engagement with sound to remedy that), there are still many things to be gleaned from the way listeners communicate about what they hear – and we hope to elaborate on that in this current undertaking. In addition, earwitness accounts typically rely on memory, rather than immediate stimulation with sound, with the exception of the practice of soundwalking, which aims at phenomenological authenticity in the listening experience. Yet even then, reflection on that soundscape happens after, and is therefore reflective and discerning on a meta-level. But what of using earwitness accounts on real-time listening? What could that immediate commentary reveal about the order in which things are heard, the significance of sonic events as they unfold in time and within the dynamic sense of context in the sound space. Where user-solicited open-ended graphic representations could sometimes be intimidating, language is familiar even if vocabulary is limited. Importantly, language is never meant to speak on its own, but offer a perspective in conjunction with soundfield measurements, audio recordings, expert characterizations or other materials.

There are several critical shortcomings of Schafer's soundscape classification system as well as of other derivative and related mapping frameworks around acoustic ecology [4, 7, 8]. As mentioned above, all of the methods are targeted at trained listeners who either report their own responses or interpret other listeners' experiences. Critics have also pointed out the inherent romaniticization of natural sound environments in Schafer's writings, in contrast to urban soundscapes which feature mechanical and electroacoustic sounds heavily. This has led to a normative hierarchy in the very classification Schafer uses to characterize soundscapes. While the idea of acoustic ecology is open-ended, the frameworks developed by Schafer and Truax are often presented as closed systems [6, 18]. These may perhaps be some of the reasons why formulations from acoustic ecology have had little to no uptake in other disciplines dealing with auditory perception and design of auditory displays. Yet we feel that a return to this unique way of conceptualizing soundscapes and listening is full of potential for understanding better how listeners perceive and interpret their auditory display-filled everyday soundscapes.

4. THE STUDY: MAPPING EVERYDAY LISTENING

As already mentioned, one of the major drawbacks to using soundscape mapping tools for the purposes of exploring the listener's perspective in an ecological manner is that these instruments are generally not validated, often exist only in prototype form and limited features prevent the representation of complex everyday soundscapes. Undertaking this project both authors build on prior work exploring listening that combines research with spatial-functional soundscape mapping through symbolic graphical notation [2]; as well as novel methodological approaches to categorizing and visualizing temporal patterns of listening/aural fluencies in the context of complex, ambient soundscapes [17]. Following a process of iterative validation, we present the first step towards developing a larger-scale comprehensive, ecological instrument for researching "everyday listening" in contexts where auditory displays play a formative role in the constitution of an *electroacoustic community*. Our project so far involves soliciting real-time listener commentary and reflection in a set of listening tasks performed with a small pilot group of participants. For this stage of the study, we have chosen two recorded soundscapes that both convey familiar everyday settings where auditory displays play a central part to form a unique and familiar electroacoustic community: one features the inside of a bank building near a set of ATM machines being used; and the other takes place at a grocery store line-up as a store clerk is "ringing" items on the cash register. We recruited 10 participants, all undergraduate students, and presented each of them individually with the two soundscapes, over headphones. Each soundscape was just over 2 minutes long. We asked participants to perform three listening tasks for each soundscape, which was correspondingly played three times in succession. In the first task, the participant is asked to identify and describe sounds that they hear in a *Think Aloud* protocol – a real-time earwitness account – as the recording plays. The recording is delivered through headphones and recorded in real-time in a multi-session track, while the participant's voice is recorded in a separate track at the same time. Upon their second hearing, they are asked to comment on the overall function of the soundscape, and after the recording is over, to discuss how well the soundscape reflected the intended function and context of the space/place. In the third task they are asked, after the recording plays completely, to create a written reflection in the form of an 'aural postcard' – a narrative about what happened in the recording, what was significant, and what sort of associations it evoked for them. The format of this study comes from a combination of Schafer's earwitness accounts, design workshop methods such as Think Aloud protocols, and ethnographic techniques such as narrativized accounts. The point is to get at several levels of phenomenological reflection on everyday listening as an experiential phenomenon – from more immediate to more conceptual/abstract. The role of analysis after then, becomes in extracting relevant patterns about the significance of listening practices in relation to the function of auditory displays within complex acoustic ecologies of everyday situations. Data collection includes integrated audio of the recorded soundscape and participant's oral account, transcripts of the oral accounts, and written reflections. Analysis includes a visual open-coding of the integrated audio, plus a more formalized stage of content and discourse analysis using the Atlas.ti qualitative coding environment that aims at identifying significant patterns of both everyday soundscapes and everyday listening. At that stage, we will be incorporating an inter-coder component in the study as well, to ensure date is consistent and reliable.

5. RESULTS AND DISCUSSION

Rather than focusing on number of correct identifications of sounds – an approach that would only reveal mechanical aural

perception – we instead shift our analytical focus on instances of specific listening approaches, and attempt to build salient patterns through careful examination of the three-tier accounts we have for each participant and each listening sample. Upon preliminary analysis, we were able to identify several emergent aspects related to the process of listening and nature of meaning-making in everyday soundscapes. Coding of the integrated audio in Tasks 1 and 2 for both soundscapes was done using the visual sound annotation tool *Sonic Visualizer* created by a development team at Queen Mary, University of London. Coding for the full study of 100 participants will be conducted using the qualitative software Atlas.ti in the form of discourse analysis. While the Sonic Visualizer tool automatically allows us to view significant events on a temporal scale, the multi-layer annotation feature allows us to juxtapose an expert's (researcher's) descriptions of the sonic events against commentary made by participants. Further, using the open-coding framework of this software we employed an iterative process involving several stages and levels of coding in order to refine a coding schema for participant responses that encompasses relevant dimensions of sonic comprehension. Based on our work so far we will discuss and illustrate four such dimensions of everyday listening – temporal, experiential, spatial and semiotic.

5.1. Temporal Dimension

Understanding how listeners hear, make sense of and shift listening modes as well as cognitive-attentional foci in a given setting is necessarily a complex process, and as much a function of perception as it is of time of exposure, level of engagement, familiarity and memory. Exploring the temporal dimension of listening in our present study consists of attempting to establish and uncover patterns in the way participants experience the given soundscape and make sense of the space, functionality and significance of what they are hearing. We are in essence looking for the temporal structure of everyday listening. Specifically, we look for stages in listening attentions as it shifts from background to foreground, or attends to sound events as opposed to sound qualities or spatial details. This temporal dimension exists in each individual task, however, taken together – the three tasks for each soundscape also add a dimension of increasing familiarity with a soundscape, and thus potential for greater reflectivity and interpretation.

In our preliminary analysis, we found that in the first task most participants tended to start by characterizing or contextualizing the soundscape – or attempting to do so; then they move on to identifying more foreground sounds, or background sound events, and in a few cases begin to associate how the sounds fit together and what sort of space, occasion or scenario is being presented to them. In the second task, overall, there is a greater level of interpretive elaboration, however, still switching back and forth between identifying potentially significant sound events, and articulating descriptive details about sounds that are heard. While in the first task it seems that listening attention is engaged with identification, in the second task the listening attention becomes more interpretive, reflective, while still tuning back in to the sound to confirm or check an assumption about the soundscape's functions and elements. Post-discussion after Task 2 and less so Task 3 (a written reflection) reveal even more reflective accounts, with more mention of the cognitive process that participants engaged in. We will return to this idea of temporal structure of everyday listening in the final discussion section again, tying together the lessons learned from the other relevant dimensions.

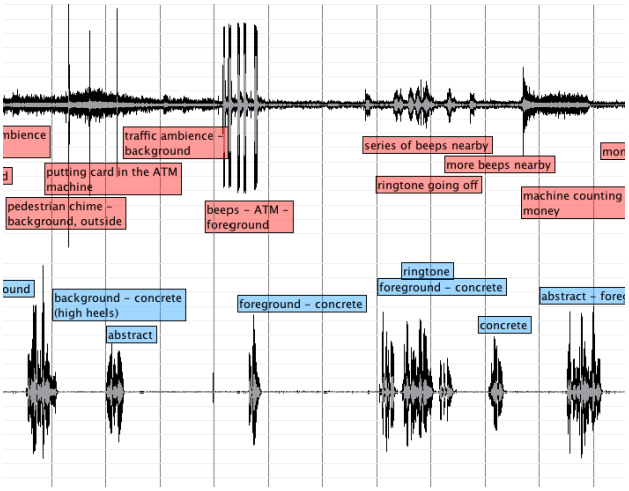


Figure 4: A zoom-in screenshot from the annotated audio transcripts from Task 1 (Soundscape1-Bank) for participant #3. Left channel – soundscape; right channel – *Think Aloud* audio; Red labels are researcher annotations of the soundscape (actual) and blue labels are coded participant comments (perceived).

5.2. Spatial Dimension

There is no doubt that sound is spatial and upon being presented with a listening task, participants are highly attuned to and responsive to the spatial and contextual characteristics of the soundscape they are hearing. This was the case in our study as well. While in the *Think Aloud* component of Tasks 1 and 2 participants often did not explicitly acknowledge whether a sound was foreground or background, in the post-discussion they relayed more detail. Again, as with the temporal dimension, the buildup of familiarity with the recorded soundscape played a central part in the attention to spatial characteristics. In Task 1, Soundscape 2 –Grocery for example, most participants correctly identified the ambience right away, even without explicitly stating how – most comments consisted of short detail about sounds in the foreground (*P9-Sounds like plastic bag noises….Canned tins and plastic noises…the products are package-based;*). Thus in Task 2 and 3, no one of the participants made specific spatial references to the soundscape – in terms of its size, configuration or depth of the various sonic signals; rather, most participants made contextual references to sounds that were familiar, which allowed them to identify the space as a grocery store and so the level of spatial observation refrained to identifying foreground versus background sounds. In other words aural comprehension shifted very quickly from contextualizing to concrete story-building of events that take place. In the Soundscape 1-Bank, most participants actually had trouble identifying the space – ideas ranged from parking lot, warehouse, factory, shop/store, office, even outside. Interestingly, in Task 2 and the post-discussion of

Task 2, many more of participants made specific and discerning references to the spatial character of the soundscape, describing in detail which aspects of the spatial character of sound led them to conclude what type of space it is: *P9 – It's a transportation station perhaps for trains or buses. Em, there's lots of echo noises around, it's a wide space, there's ongoing construction and there's um the moving metal trollies. And you can hear the echo noise, em, the long reverberation or echo noises of transport nearby.*

5.3. Semiotic Dimension

The semiotic dimension of this everyday listening exercise reflects the informational, associational and general 'sense-making' strategies that participants engaged in trying to understand the two soundscapes. Utilizing an open-coding iterative approach to participants' comments in all three tasks we devised a classification system for the way participants described and identified sounds, resulting in several more granular categories: **Sound Typologies: concrete vs. abstract** sound references**; Associational** sound identifications**;** and **Narrative** elaborations (see Table 1). Naturally, most often, each listening experience or instance of listening entails a combination of these approaches.

5.3.1. Sound Typologies: Concrete / Abstract identification

Concrete references involve mention of particular sonic objects, events or situations, while abstract sounds merely refer to the general sonic character or sound quality of what is heard. Concrete references by participants in Task 1 included comments such as "a beep", "woman speaking", "footsteps", while abstract references included "a shuffling", "loud noise", "high-pitched sound". The difference, essentially, is one of degree or level of identification of a sound even in a general way, as opposed to a reference only to the general idea/character of the sound in more abstract terms without necessarily specifying it. Sound events and details could be said to be a type of concrete sonic reference that go further than a concrete acknowledgement and refer to implied action or physical-interactional properties of the object. Sound events are indicated by participant comments such as "Things being dropped.", "Cages opening and closing", "Trolley being pushed…keys being pressed." Sound details include more direct references to the materials and interactions of sound such as "metal cages", "tin cans", "rustling of plastic bags", "package-based items". As mentioned above most often participant comments involved a combination of several levels of sound identification. To exemplify, we look in detail at the Task 1 transcript of Participant 6, listening to Soundscape 1-Bank: it starts with four foreground beeps, identified by the participant with a *concrete* reference; the soundscape continues with some mid-ground beeps and a very short mobile phone ring in the distance, identified by the participant as a *concrete sound event* of a Nokia phone; this is followed by the foreground sounds of an ATM accepting a card in the slot, then counting money and dispensing them – identified by the participant as the *concrete* sound "of a cash machine", accompanied by an interpretive gander at the meaning – "a ticket machine printing maybe" – as an *associational* reference. This type of meta-level coding allows us to get beyond individual reporting styles and look at more general patterns across participants in their semiotic approach auditory information in an everyday context.

5.3.1. Associational sound identification

Associational references were the most common of commentary for both soundscapes on all three tasks. Associational references entail an explicit or implied association to a familiar, past experience or sound, resulting in a cognitive synthesis between what is heard and what it 'sounds like' – a type of template-matching. The way we identify those is that most often participants will preface a reference to a sound or event by saying "It sounds like…" which is typically always followed by an interpretive statement – "It sounds like a tape being put in, a tape recorder or machine of some sort", in contrast to more direct identification such as "a beep", "a machine sound", "another beep". Associational cues are key to understanding how participants make sense of a complex, everyday soundscape semiotically, and is particularly important to the identification of auditory displays as many of them are quite similar in tonal character, thus resisting a clear 'auditory template'. Association – which entails familiarity and drawing on prior experience with similar sounds seems to be, even in our small study, overwhelmingly the main technique that participants employ in listening to these soundscapes. Both soundscapes were rich in simple auditory displays – beeps and related signals – strikingly similar in tone/duration/quality even as the contexts were completely different. Perhaps it is that generic similarity that drove participants to rely largely on associative and contextual cues. Curiously, the only two sounds that were explicitly and correctly identified by all participants were the mobile phone ring in the background of the bank soundscape, and the one error beep on the cashier till at the end of a busy transaction in the grocery store soundscape. Clearly, given that we listen for difference and adapt to similarity, it was those two out-of-place sounds that attracted attention and seemed important enough for participants to report on.

5.3.1. Narrative Elaborations

Narrative references involve a higher level of association in the form of what we'd call *imaginative listening*. While associational cues generally consist of interpretation on a single or discreet sound event, narrative references entail entire scenarios – stringing together sonic cues into a coherent story, narrating the events that are [potentially] taking place, and in that, referencing contextual details that are not in the original soundscapes. In the case of Soundscape 1-Bank, narrative accounts did not surface until the post-Task 2 discussion (*P4 [who thought this was a car park underground] - somebody's phone going off, somebody's phoning them to find out where they are of if they, you know, just parked the car, and they're just getting out of the car*). Since most participants did not correctly identify Soundscape 1-Bank, but did correctly identify Soundscape 2-Grocery, associational cues in conjunction with narrative constructions reveal a lot about the process of listeners' meaning-making. For those in Task 1 who thought the bank environment was a car park, every beep became "the sound of vehicle reversing", while the rumbling of the ATM counting money and dispensing them became "motor or

machine sounds" or "engine starting". For those who interpreted the soundscape as an office, beeps became "sounds of scanners or equipment" and the close-up ATM mechanical sounds became "a photocopier, someone pulling out paper". Participants who did more free-association on the first task and referenced a warehouse, a photocopier and a tape deck at the same space, commented on the incongruence of those sound signals in the discussion after Task 2 as they didn't quite fit into the *story* of that space. In Soundscape 2-Grocery, conversely, as early as Task 1 many participants narrated rather than identified sounds – they narrated the exchanges and almost visualized the events taking place (See some examples in Table 1.). Some participants even imagined inaudible events ("customer is probably passing off a club card of some sort"), others reported on how many tills there might be ("small shop – around 3 tills") or how many customers were present in general ("heard about 5 customers"). In the subsequent tasks for this soundscape, participants had no trouble integrating all the sounds they heard as belonging to a space that they immediately identified with a supermarket. Beeps didn't signify machinery here, but rather evoked deeply human exchanges, the "general hustle and bustle of a supermarket".

5.4. Experiential Listening

Experiential listening we'd put in its own category in order to capture instances where participants referred only to sound parameters and subjective listening characteristics such as loudness, pitch, sound colour; including their use of onomatopoeia words to identify and references sounds. While experiential listening is probably the most primary of impressions phenomenologically speaking, as far as the task sequence were concerned it tended to come up in more reflective discussions, higher level analysis, rather than in first-person narration. It seemed to be engaged more – similarly to spatial listening – when the soundscape is perceived to be more unclear, ambiguous in terms of purpose and setting.

6. SUMMARY AND DISCUSSION

What we aimed to do at this preliminary stage of the study is identify the temporal progression or structure of listening to auditory displays within everyday soundscapes that entails dynamic shifting of listening from contextualizing, to identifying, to associating, to spatially locating and interpreting sonic signals. A pattern in that temporal progression might help us understand how listening functions over time and thus *design* for it better – particularly in contexts of more ambient, multi-lateral soundscapes or in cases of more complex auditory displays. From a research perspective, this helps us identify more comprehensive tools for soundscape mapping that takes into consideration the temporal and contextual dimensions of everyday listening. In Table 1 below we synthesize the elements of soundscape perception and comprehension that has emerged from this pilot as an ontology of everyday listening for the purposes of coding and analysis of our larger study sample of 100 participants. Through an iterative process we have herby distilled useful definitions that we propose are general enough to be usable to other research explorations oriented towards the contextual and temporal nature of listening to auditory displays in ecological settings.

Spatial	Making specific references to space including proximity, size, architectural features, etc.: Inside/outside; close/far; big/small space; echo/reverb; left/right/up/down ***Example/Instances of Use:*** *P5 -I think it was either a factory or an office, but I think it was actually a bigger space than an office, or it might be a corridor in a office, but I, more I think it was actually a factory…because of all the echoes around…*
Experiential	Describing the quality of sounds as they are experienced; use of onomatopoeia words; reference to any sound parameters: Loud/quiet; timbre, pitch, rhythm, etc. ***Example/Instances of Use:*** *P4 – Very reverberant…the people's feet on the floor was quite a hard sound, like heels hitting concrete*
Semiotic	
Sound typology	**Concrete/abstract:** identifying and naming specific sounds/ identifying only general character of sounds; typically refers to sound event/action (not source). ***Example/Instances of Use:*** *P5 – Cars. Beeping of a machine. More beeping. Footsteps. Switches being pressed. More beeps. Mobile phone.*
Association	Free association based on what is heard; using associative language; limited to references to 1-2 single/individual sounds. ***Example/Instances of Use:*** *P8 – Sounds like a trolley, being wheeled around. P2 – Sounds like a bus of some sort, a vehicle taking off. And then some beeping, which could be, a vehicle reversing or something*
Narrative	Connecting several (2+) sounds together to build a story of what happened; interpreting a combination of sounds to put a sequence of events together; a higher level of associational thinking. ***Example/Instances of Use:*** *P1- Em, next customer's coming along, put their stuff through the scanner, again you can hear the beeps of the scanner…the customer just said they had a bag so I'm assuming the cashier's offered them one…Em, can hear what sounds like stuff being taken and placed into a bag, the rustling of a plastic bag*

Table 1: A schematic breakdown of the elements we identified in the temporal structure of everyday listening to complex soundscapes that feature auditory displays.

To summarize our preliminary study results, taking into account all the dimensions discussed so far, we suggest a guiding schema that reflects the listening and sense-making process that people generally follow in a soundscape listening task. As shown in Figure 5, everyday listening entails first an attention to the context, situating the listening experience; then a focusing on sound events, switching attention between foreground and background sounds and focusing on concrete identification; and ultimately associating – combining what is heard to what is known about the context and the memories of similar experiences, attempting to make coherent narrative of the experience by linking and integrating both present and associational material.

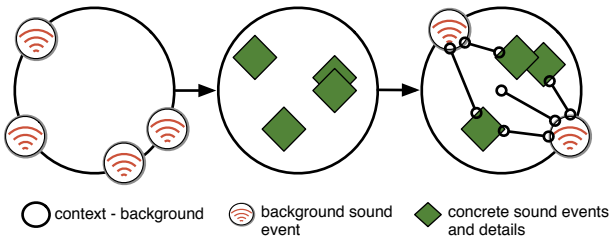


Figure 5: A conceptual model illustrating the process of listening our participants engaged in – from contextualizing sounds to identifying and interpreting them, to putting them in a coherent story.

7. CONCLUSION

To return to our initial impetus for the study - developing a research tools for exploring everyday listening that incorporates not only perceptual and functional but also ecological and contextual dimensions, there is admittedly still much work to be done. In the next stage of this analytical process, we plan to code data from the full study of 100 participants with the finalized coding schema presented here, then draw out analytical and quantitative measures towards a conceptual synthesis of listening to auditory displays in everyday settings. The full content analysis of all task transcripts should allow us to reinforce some of the conclusions proposed thus far regarding the temporal structure and sequence of listening comprehension.

The main contribution that we feel this work makes to the auditory display community is in offering a framework for incorporating acoustic ecology aspects into the validation and use of research instruments aimed at understanding and examining how people listen to auditory displays in everyday sound settings. This study puts forth a sophisticated analysis of listening in temporality bringing experiential impressions together with cognitive processes in real time. By analyzing listening modes/attentions in this way we can see what is being prioritized, what is focused on, what is lost. Even at this preliminary stage, we are able to offer a guiding structure of relevant dimensions that focus on facets of listening not typically represented in other instruments for soundscape mapping, listening task studies or field testing of auditory displays. The associational nature of listening and its importance to the contextualization, correct identification and construction of meaning with regard to auditory displays in a given soundscape is something not typically reflected in traditional perception research. Further, the lack of validated instruments for qualitative research of listening; including the use of sound maps is a gap in need of further work. It is in those areas that we situate our work and hope to make a contribution to, enriching the field of auditory displays with more interdisciplinary theory, methods and approaches. As auditory displays increasingly build into social memory and become perceptually drawn upon by listeners in everyday environments, researchers have no choice but to consider more ecological approaches to understanding perception and auditory cognition.

8. REFERENCES

[1] McGregor, I., Leplâtre, G., Turner, P. and T. Flint, (2010) Soundscape Mapping: A Tool for Evaluating Sounds and Auditory Environments, *In Proceedings of the 16th International Conference on Auditory Display,* pp. 237-244.

[2] McGregor, I.; Leplatre, G.; Crerar, A.; Benyon, D. (2006) Sound and soundscape classification: establishing key auditory dimensions and their relative importance, *In Proceedings of the 12th International Conference on Auditory Display,* pp.105-112,

[3] Gaver, W. W. (1993). What in the World do we Hear? Ecological Psychology, 5(1), 1-29.

[4] Hellström, B. (1998). The Voice of Place: A Case-study of the Soundscape of the City Quarter of Klara, Stockholm. In R. M. Schafer & H. Jarviluoma (Eds.), Yearbook of Soundscape Studies 'Northern Soundscapes', Vol. 1, 1998 (Vol. 1, pp. 25-42). Tampere: University of Tampere, Department of Folk Tradition.

[5] Coleman, G. W., Macaulay, C., & Newell, A. F. (2008). Sonic mapping: towards engaging the user in the design of sound for computerized artifacts 5th Nordic conference on Human-computer interaction: building bridges (pp. 83- 92). Lund, Sweden: ACM.

[6] Schafer, R. M. (1977) *The Tuning of the World*. New York: Knopf. Reprinted as Our Sonic Environment and the Soundscape: The Tuning of the World. Rochester, VT: Inner Traditions International, 1993.

[7] Southworth, M. (1969). The Sonic Environment of Cities. Environment and Behaviour, 1(1), 49-70.

[8] Schiewe, J., & Kornfeld, A.-L. (2009). Framework and Potential Implementations of Urban Sound Cartography 12th AGILE International Conference on Geographic Information Science.

[9] Heckl, M., & Müller, H. A. (1994). Taschenbuch der Technischen Akustik. Berlin: Springer Verlag.

[10] Zwicker, E., & Fastl, H. (1999). Psychoacoustics: Facts and Models (2nd ed.). Berlin: Springer.

[11] Valle, A., Lombardo, V., & Schirosa, M. (2009). A Graph-based System for the Dynamic Generation of Soundscapes. In M. Aramaki, R. Kronland-Martinet, S. Ystad & K. Jensen (Eds.), Proceedings of the 15th International Conference on Auditory Display. Copenhagen, Denmark: ICAD.

[12] Augoyard, J.F. & H. Torgue (2005) Sonic Experience: A Guide to Everyday Sounds. Montreal, CA: McGill Queen's University Press.

[13] Torigoe, K. (2002). A City Traced by Soundscape. In H. Jarviluoma & G. Wagstaff (Eds.), Soundscape Studies and Methods (pp. 39 - 57). Helsinki: Finnish Society for Ethnomusicology; Department of Art, Music and Literature.

[14] Hedfors, P. (2003). Site Soundscapes: landscape architecture in the light of sound. Unpublished Ph.D., Swedish University of Agricultural Sciences, Uppsala.

[15] Giaccardi, E., Eden, H., & Fischer, G. (2006). The Silence of the Lands. Proceedings of the New Heritage Forum, 94-114.

[16] Stratoudakis, C., & Papadimitriou, K. (2007). A Dynamic Interface for the Audio- Visual Reconstruction of Soundscape, Based on the Mapping of its Properties Proceedings SMC'07, 4th Sound and Music Computing Conference, 185-191.

[17] Droumeva, M. & R. Wakkary (2010) Socio-ec(h)o: Focus, Listening and Collaboration in the Experience of Ambient Intelligent Environments, *In Proceedings of the 16th International Conference on Auditory Display*, pp. 327-334.

[18] Truax, B. (2001) Acoustic Communication. 2nd Ed. Norwood, NJ: Ablex Publishing.

[19] Andean, J., Akatemia, S., Järviluoma, H.; Kytö, M.; Truax, B.; Uimonen, H.; Vikman, N. and R. Murray Schafer (2010) *Acoustic Environments in Change & Five Village Soundscapes.* TAMK University of Applied Sciences, 431 pp.

SONIFICATION OF PRESSURE CHANGES IN SWIMMING FOR ANALYSIS AND OPTIMIZATION

Thomas Hermann[1], *Bodo Ungerechts*[2], *Huub Toussaint*[3], *Marius Grote*[2]

[1] Ambient Intelligence Group, CITEC, Bielefeld University
[2] Neurocognition and Action Group, Faculty of Psychology and Sport Science, Bielefeld University
[3] Human Movement Science Group, VU University of Amsterdam
thermann@techfak.uni-bielefeld.de

ABSTRACT

This paper introduces the sonification of pressure sensor data measured while executing crawl stroke swimming. Swimming research aims at better understanding the flow conditions in detail to adapt swimming strokes to achieve maximal speed with minimal energy consumption. The fact that a pressure field is induced during the interaction of body and water is rarely considered. Any aquatic self-induced locomotion needs a *mediator* to cause a reaction in terms of body motion since there is no solid object a swimmer can push off from. The mediator function is taken over by the pressure field caused by the swimmer's actions. With our sonifications of the mediating hydrodynamic pressure – measured at 5 positions along one arm – we turn the hydrodynamic situation into a complex sonic rhythmical motive. These motives become auditory gestalts and we can identify differences and variations between patterns. We present six alternative sonification methods and discuss the resulting sounds in their ability to bring different patterns to attention. Our future goal is to help swimmers to optimize their motions by real-time sonification.

1. INTRODUCTION

Sonification allows to combine multiple data channels into a single sound stream, enabling listeners to understand coherences in the data that could otherwise be overseen. Similar to our ability to perceive simultaneously playing orchestra instruments as a musical piece, yet also to focus on a single instrument, we can benefit from multi-stream sonifications on different levels, such as for process monitoring, data analysis or diagnosis.

A particularly promising application field is the use of sonification to understand and support the coordinated movements of the human body, e. g. in dance, while playing a musical instruments, or during sports such as rowing [1], swimming [2], speed skating [3] or German wheel training [4].

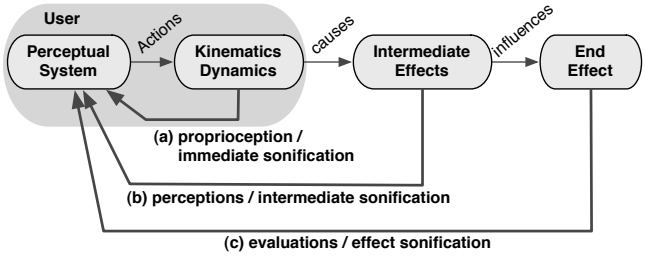


Figure 1: Sonification as Auditory bio-feedback: sonification can provide information from (a) the immediate state, (b) the intermediate effects, and (c) effect information. We suggest that the intermediate level may offer valuable information as a scaffold for learning.

The main benefits of sonification in the area of swimming research discussed in the paper are that (i) sound is accessible without demanding visual attention (which would be difficult underwater), that (ii) our auditory perception has a high temporal resolution, allowing tightly closed interaction loops in online applications, and (iii) we are highly sensitive to rhythms and changes of rhythms, and these patterns occur frequently in repetitive coordinated body movements.

1.1. Sonification of Intermediate Levels

Most sonification approaches in movement research research start from body postures and sonify the kinematic information to understand or support the execution of movements (few selected references are [2, 5, 6, 7]). On the other side, there are sonifications that represent the overall task-specific effect (such as the intracyclic fluctuation velocity in rowing [1]) as the source for sonification. Both feedback types enhance the better perception of the users' actions and their effects. However, we suggest that complex goal-driven actions can be regarded as a chain (as illustrated in Fig. 1) or even better as a continuum that have *intermediate* processes between the users' actions and the ultimate task-specific

effects. Intermediate processes are all physical processes between the actions and their intended effect. Direct feedback on the behavior (e.g. kinematics, or the deviation from a nominal movement) may help to induce a specific motion pattern, yet this will not necessarily guarantee the wished total effect. On the other side, a mere feedback of the effect variable (e. g. the overall speed) may lack suitable information for the users how to refine their motion to achieve better results. Here we suggest to sonify the *intermediate effects* which are caused by the actions and in turn influence the end effect. It might indeed be difficult or impossible for a user to integrate an intermediate feedback for the self-regulation in the actual movement execution, but if the movement is a repetitive pattern, the user might be able to explore how the own actions systematically relate to sound changes and refine the movements for the subsequent repetitions.

This paper takes swimming sonification as an example for intermediate effect sonification. In contrast to former sonifications of swimming actions focusing on the distance of hands from the body [2], in this paper sonification represents the intermediate effect of hand actions that displace water and thus induce flow pressure. Specifically, we focus on flow patterns in sport swimming.

1.2. Sonification of swimming

According to elite swimmers' saying, effective swimming is a matter of "feel for motion of water mass" controlling the interaction of water mass and body limbs. However, little is known how to communicate this kinesthetic wisdom. Mostly swimming actions are studied via the kinematics of the external gestalt, leaving out the motion of water mass. Motion of water mass, however, induces hydrodynamic pressure and together with the pressure of the water column the entire interaction is represented by measuring the total pressure. The transformation of pressure signals into force-time-data may inhibit information because force is finally not a kinesthetic valuable, e.g. muscle tension. Our starting hypothesis is that the sonification of pressure offers a helpful new channel to support the communication about flow and on the sensation of flow, respectively.

Our primary goal in this paper is to develop and introduce sonification methods that allow to investigate the patterns of total pressure that occur during crawl stroke swimming using previously recorded data sets and videos. Thus we offer different methods to make patterns accessible as sound, we sonify data from different crawl speeds, and listen to the sounds to characterize the sonification methods in their ability to uncover relevant structures. In a future step these methods may serve as the basis for future online sonifications to be developed as a new teaching and training instrument, also to be used by swimmers in the water for self-regulation. Practically this can be done for instance by using underwater loudspeakers[1] or available swimming solutions for earphones. The latter do not only enable stereo sound projection, they also reduce the level of external sounds such as water splashing [2]. For sound rendering a belt-mounted mobile phone with underwater protection may be used.

The paper starts with an introduction to swimming research followed by some explanations of the relevant phenomena and the origin of the data. Section 3 introduces the selected data sets and explains the features to be used for the sonifications. Section 4 presents six sonification methods, explains why and how they have been selected and illustrates them with sound examples. We then discuss what auditory patterns stand out or surprise. The paper concludes with a discussion of the results and an outlook on future work.

2. SWIMMING RESEARCH – APPLIED HYDRODYNAMICS

This paper is about the sonification of water, set in motion by hand actions during crawl stroking. Whatever is said about the aquatic effect of hand actions, the origin of propulsion is still a matter of discussion. In most cases the kinematic aspects of body actions are emphasized. The hand action during crawl stroking is a cyclic 3D event in aquatic space and can be described using functional analysis whereby the following nodes (BACs)[3] are used: (1) Fingers enter water, (2) Hand moves forwards, (3) Body rolls to side of action, (4) Hand moves downwards, (5) Prolonged pronation of the hand, (6) Hand moves upwards, (7) Body rolls back, (8) Slicing hand moves outward, (9) Breathing in, (10) Hand moves forward. The duration of action below waterline is approx. 70–85% and above is 15–30% per cycle. In most cases the kinematic aspects of these body actions are emphasized. However, without regarding the interaction between hand and water mass the story is incomplete like the description of applauding with one hand. In the field of biomechanics of swimming, the conditions of self-induced locomotion is still a matter of discussion. Traditionalists emphasize the application of steady flow physics as used e.g. in ship construction. But the effects of hand actions cannot be limited to a question of forces since forces do not explain their origin.

Meanwhile the change of body form (per cycle) and the creation of unsteady flow conditions are recognized as a central aspect. Unsteady flow in the vicinity of a body is characterized by changes of flow velocities in time and space. In particular three agents are involved to generate effects: the body (i.e. the propelling parts like hands and feet) moves water mass while a pressure field is induced.

[1] e.g. Ocean Engineering Enterprises "OCEANEARS" (DRS-8)

[2] e.g. http://www.h2oaudio.com/store/flex-waterproof-all-sport-buds-super-hero-blue.html

[3] BAC = basic action concepts, see [8]

The term pressure in flowing water can be distinguished into hydrostatic, static and hydrodynamic pressure. Hydrostatic pressure depends on the mass of water and column height which represents the potential energy. Static pressure is like normal stress elementary to particles (mass and volume) at rest or streaming with others due to exerting pressure in all directions (like compression). Hydrodynamic pressure is an induced component due to the (local) flow velocity, representing kinetic energy. The sum of all is called total pressure. When the water is displaced by a body the total pressure is of particular interest. Displacement of streaming particles demands some pressure work (on a certain volume of water) which is an amount of work to force some mass m of a volume V from a certain pressure p_0 in a space with the pressure p_1. Due to this, the mass transfers some of the potential energy into kinetic energy by means of a third energy, the pressure work. Those locally altered pressure components induce "proto-vortices" [14] which contribute to locomotion whereas pressure drag is not a major player. Hand motion, starting from the water line where the hydrostatic pressure is small, is directed to a deeper level accompanied by increasing total pressure and finishes at the water level again. Continuous displacement of water mass by the hand induces a change of hydrodynamic pressure.

The secret to maximally propel the body forward per cycle is to move the hand continuously along a curved 3D line shaped like a crescent bowl, starting from (BAC 3) until (BAC 8). This movement makes use of the change of potential energy to kinetic energy by means of pressure work. This is what makes swimming so exhausting, except when a jet stream is created due to vortex-like flow structures, as they occur in tornados as a matter of the pressure distribution.

Swimming research is documented by a series of "International Symposium of Biomechanics and Medicine in Swimming" organized every four years since 1970. During these decades several studies related to pressure measurements were presented as well. Van Manen et al. [13] expect that wrong hand positions can be explained when unusual pressure graphs occur. Takagi and Wilson (1999) [11] put forward that without pressure no propelling force will be produced and a pressure differential method is potentially a useful means in stroke analysis.

Toussaint et al. (2002) [9] studied the pressure along the extremity of elite swimmers executing crawl stroke to investigate the axial flow component. Waterproof pressure sensors have been attached to different body point (shoulder, elbow, wrist, dorsal and palmar side of one hand, see Fig. 2) and calibration was done by measuring the hydrostatic pressure at different depth in water. Total pressure signals were recorded and low-pass filtered at 25 Hz while swimming at slow, intermediate and sprint speed. The key assumption is that flow effects act predominantly perpendicular to the local measuring point. Comparing total pressure–time-curves of

Figure 2: Sensor setup used to measure the pressure data at hand palm, hand back, elbow and shoulder.

all measuring points at sprint speed globally they show individual shapes and data were highest at the palm, at the dorsal side of the hand and at the elbow approx. 60% less, and at the shoulder lowest, approx. 80% less relative to palmar pressure, before all curves descend and turned remarkably to suction during the last 1/3 of the cycle period. Since dorsal pressure drops much more, the hand does not act like a paddle. When the pressure at the dorsal side of the hand is lower than the pressure at the shoulder this is completely opposite to what is hypothesized when taking the effective water column into consideration: the hand is deeper than the shoulder). A local pressure drop near the fingertips will induce an axial fluid flow along the arm and hand towards the fingertips which lead to an increased propulsion (pumped-up propulsion) and it suggests that swimming faster is more a matter of decreasing the pressure at the dorsal side of the hand than augmenting the palmar pressure. How these results can be used in practical questions such as teaching or self-regulation needs still to be evaluated.

Loetz et al. [12] point out that pressure-time recordings are an "essential complementary information". In search of communicating this information the sonification of pressure data might be a promising tool, not only because pressure waves and sound waves are alike. Since the link between kinematics of the hand and the resulting pressure or propulsion is not fully understood, a better communication between swimmers/experts is needed. Our vision is to give feedback to the swimmer directly – probably in conjunction with an effect variable such as the intracyclic velocity-variation – and to support the communication about flow and the sensation of flow between all experts. A necessary first step is to examine how sonification can be used for making a pressure field audible. For this first step the data of an experimental study published in a peer-reviewed journal by Toussaint et

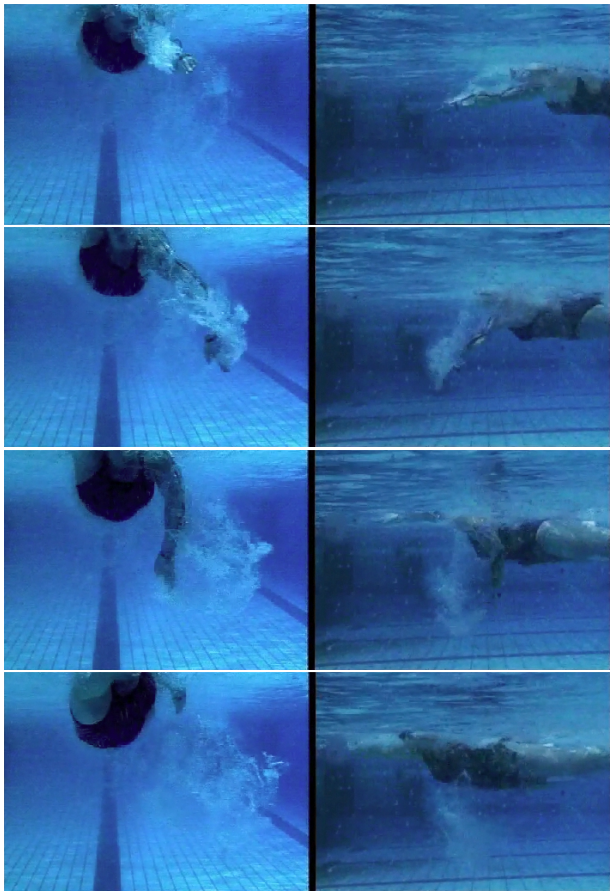


Figure 3: Key frames of a crawl stroke from front (left frame) and side (right frame): the air bubbles allow to understand the 3D trajectory. The video was recorded by the 3rd author and corresponds to the condition 'faster' shown in Fig. 4.

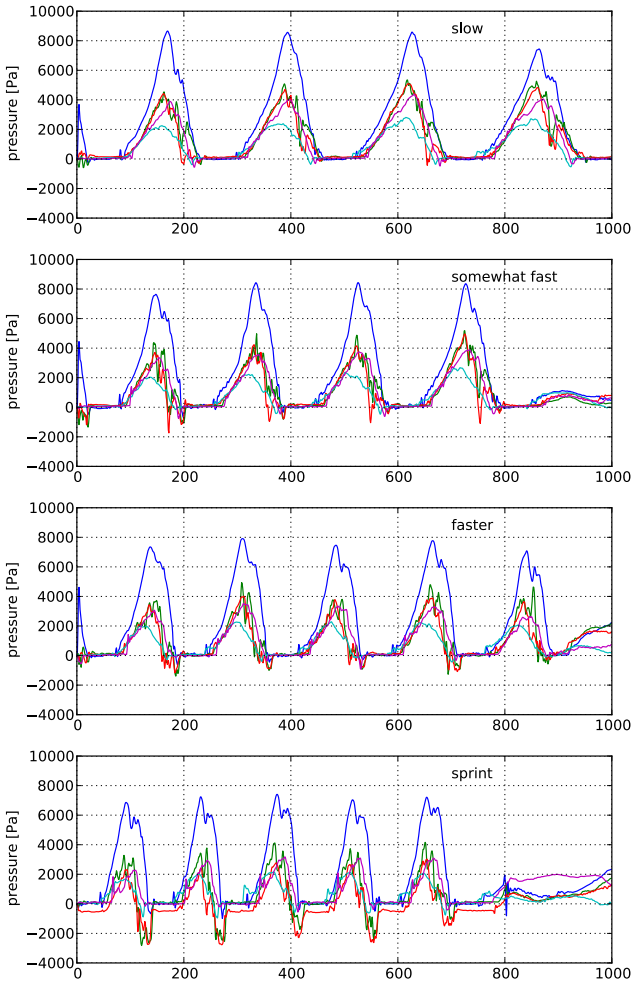


Figure 4: Pressure data at selected points of the one arm: shoulder (cyan), elbow (magenta), 1/3-elbow(red), palm of hand (blue), back of hand (green) as function of time for different crawl-velocities. The data are recorded at 1000 Hz, filtered to 25 Hz and down-sampled to 100 Hz.

al. in 2002 [9] were used in this paper.

3. DATA AND FEATURES EXTRACTION FOR SONIFICATION DEVELOPMENT

For the development of the sonification methods we start with pre-recorded sensor data measured at different points along the upper limb of an elite swimmer in a study done by the co-author [9]. Fig. 2 shows the sensor setup attached to the arm of the swimmer. Selected video key frames of a crawl-stroke in the data set 'faster' are depicted in Fig. 3. Fig. 4 depicts the data sets for 4–5 crawl-strokes at slow, somewhat faster, faster, and sprint performance. The flat plateau between the strokes around a pressure of 0 Pa represent the intervals where the hand has left the water. While visual inspection allows to discover certain patterns such as the acceleration of the rhythm or the decrease of pressure below 0 Pa for the back of the hand at sprint, it is more difficult to understand temporal patterns that involve all 5 time series from visual inspection alone.

We started from basic direct sonifications and gradually advanced towards task- and analysis-specific auditory displays that render features more salient that are expected to be relevant for understanding the phenomena. In this section we summarize data features and their computation as they are needed in the following section to specify the mappings.

Polarity: Firstly, we see that the data is ordinal with a defined zero value. To better perceive the polarity of a time series, it makes sense to use a feature $f_p(t) = \text{sgn}(x(t))$. This feature, however, would exhibit many value changes when the pressure oscillates around 0 Pa so that a modified

feature is superior which returns 0 if the value is below a threshold θ_0. A suitable value is around $\theta_0 = 150\,\mathrm{Pa}$.

Slope: The gradient can be computed by

$$f_g(t) = \nabla x(t) \approx (x(t) - x(t-\tau))/\tau \qquad (1)$$

where τ is 1/sampling rate. Since the data are low-pass filtered, this feature is quite stable and will be used for excitatory sonifications.

Local Maxima/Minima: for event-based sonification, local optima as well as zero crossings are candidate time points. Since the time series is low-pass filtered, a 3-point criterion provides a suitable condition to detect extrema:

$$\begin{aligned} f_{\min}(t) &= (x(t-\tau) > x(t)) \wedge (x(t) < x(t+\tau)) \\ f_{\max}(t) &= (x(t-\tau) < x(t)) \wedge (x(t) > x(t+\tau)) \end{aligned}$$

4. SONIFICATION METHODS

A data set is basically a 5-dimensional time series and there are manifold possibilities to sonify them, starting from a naive time-variant frequency modulation to task-specific designs. We document the development cycle and report six selected sonifications that provide gradually different 'sonic views' of the data. Please note that in this first design stage we are primarily interested in the sort of sound patterns that emerge when sonifying the data – we do not consider the aesthetics or the compatibility with environmental sounds here, yet we acknowledge that for any practical applications these are major factors for subsequent optimization. All approaches demand the manual selection of parameters (e.g. frequency ranges, level ranges, etc.). Most of them have been subjectively adjusted, and thus depend on personal design experience and taste. Limited space prohibits to discuss all choices in detail. Certainly, such parameters are subject to swimmer-specific personalization, should a method be selected for further consideration. Since we consider the sonifications as preparation for future real-time/online use, we map the real time to the sonification time throughout all methods. For detailed analysis, however, we provide 1:3 slowed down sonifications.

4.1. Standard oscillator bank mapping

As the data is in essence a multivariate time series, the first approach was to sonify the data in the most direct and naive way, using a simple mapping of the values to a bank of 5 sine oscillators. This provides a rough first sketch of the dynamics that is to be expected from the sonification. The mapping spreads the channels equally in spectrum, from upwards from shoulder, elbow, 1/3-elbow, via hand back to hand palm, one octave per channel. The pitch range is 9 semitones, ranging from the minimum to maximum values in the time series. Listening to sonification examples (see website[4]) S1a (slow), via S1b (somewhat fast), S1c (faster), S1d (sprint) allows to perceive the rhythm and the speed. Interestingly a different timbre is audible at the 'zero-pressure breaks' where the hand is above the water. This is because the mapping maps the min/max pressure range to the min/max pitch range, causing different pitch values for the zero-pressure values. The increasing pitch indicates that negative pressure (suction) increases on average with crawl-speed. An interesting pattern is, that the higher pitched tone leads (or preceeds) the change in the pitch wave. This pattern becomes even more salient in the following sonifications. Finally sound example S1e is a 1/3 slow-motion sonification of the first two crawl-strokes of the sprint data. We find that this slower pace makes it much easier to attend to patterns for analysis and learning, yet we think that with increasing familiarity with the features, real-time interactive use will be feasible.

4.2. Excitatory Oscillator Mapping

The naive mapping has the disadvantage that the sound remains equally audible independent of the activity. Therefore in this approach we create a sonification that remains soft to inaudible when the signals are constant. Practically, this is achieved by mapping the absolute value of the derivative $|f_g(t)|$ of each time series to the level of a white noise signal which is fed into a subtractive synthesis with controllable ring time and center frequency. Pitch depends on the value just as before, so low-pitched sounds correspond to the shoulder, high-pitched sounds to the hand. Yet now the polarity of the signal is additionally mapped to the spatial panning. In result negative pressures (which are here of particular interest) become salient as they are represented by sounds from the left audio channel.

The sound examples S2a, S2b, S2c, S2d are sonifications for the different speeds (slow, somewhat faster, faster, sprint). The emphasis of change makes activity audible and particularly it can be heard that a high-pitched action preceeds the larger sound wave. For faster speeds, it becomes audible that there is a distinct pitch curve at the end of each crawl-stroke, related to the negative pressures. It sounds like the high-pitch actions (hand) 'frame' the overall stroke. This becomes even better audible in the 1/3-slow motion sound example S2e.

4.3. Single-stream multi-parameter mapping

Multi-parameter mapping is an approach that binds different channels more tightly together into holistic perceptual units than the above multi-stream approaches. The time series is mapped to different parameters of a *single* continuous sound stream. The only problem is to find a good motivation

[4] see http://www.techfak.uni-bielefeld.de/ags/ami/publications/HUTG2012-SOP

for the specific selection of which time series controls what parameter, which may appear quite arbitrary. Yet once the mapping is defined and kept constant, it may just be learnt by heart and understood implicitly and then the sounds may be useful nonetheless. Specifically, we used a formant filter synthesis with pitch, level, center frequency, bandwidth, and panning as the 5 different parameters. The detailed mapping is as follows:

hand back	[min, max]	→	freq	[80 Hz, 120 Hz]
hand palm	[min, max]	→	cf	[200 Hz, 800 Hz)
1/3 elbow	[min, max]	→	bandwidth	[100 Hz, 1000 Hz]
shoulder	[min, max]	→	panning	['right', 'left']
elbow	[min, max]	→	level	[-40 dB, -6 dB]

We received a first opinion from the swimmer whose data has been recorded for the sonification who felt that the sound reminded her of a 'tortured cat'. Clearly such issues need to be considered once a design is to be optimized for sustained use. Concerning the patterns, the sounds allow the listener to follow the roughness of the wave around its maximum, and it becomes audible that there is an increasing roughness (in brightness and pitch) at the main wave with increasing crawling speed.

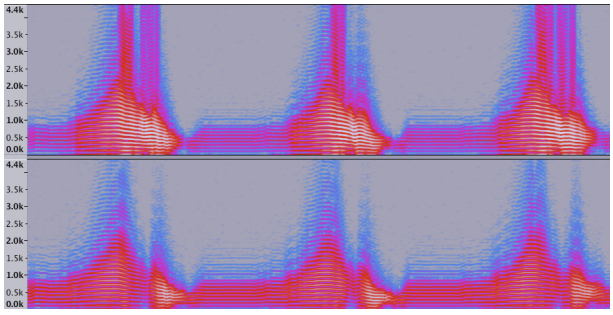


Figure 5: Spectrogram of the single-stream sonification: upper plot shows the left stereo channel. 3 strokes and subtle changes in level, brightness, and panning can be observed. This plot depicts the beginning of sound example S4a.

4.4. Harmonic Series mapping

Timbre is a multidimensional parameter, and while timbre itself may be difficult to characterize and memorize, timbre changes can be quite salient and characteristic. This motivates a variation of the previously demonstrated single-stream approach where now an additive model is used so that the different pressure variables control the activation of different harmonics. In result, the *timbre* – characterized by the amplitudes in the harmonic series – changes according to the pressure in the channels. A continuous playback, however, causes the harmonics to separate into different sound streams. For that reason we added an LF pulse to chop the signal into segments. Thereby we get a coherent onset in all harmonics which enhances timbre perception and differentiation. The pulse rate itself is a very salient parameter, and here it is used to represent the total pressure, while the hand back pressure is mapped to the fundamental frequency, but using only a small pitch variation, so that the timbre change achieves a balanced saliency.

Sound examples S4a–S4d are sonification for the different speeds from slow to sprint. S4e is, as above, the 1/3 slow-motion sonification. The sound supports the observation made above that activity in some channels (here: higher harmonics, hand) frame the major pressure wave.

4.5. Event-based Mapping

While all previous approaches started from a continuous representation of the time series, this approach follows the idea that continuous sonic information may deliver overly detailed information – in fact a condensation of the detailed values to 'key frames' of the pressure curve may not only leave the sonic signal easier to process, but we expect that this makes slight differences in synchronization between the different channels much better perceptible since they lead to changing patterns in the sequence of events. Practically we consider zero crossings (in both directions) and minima/maxima as the most relevant event types. For both minima and maxima, the actual value and the level value to the previous extremum of the other type are variables that can be used to parameterize details of the events. Sonification examples S5a–S5d start with the representation of zero crossings. The slope at the zero crossings is mapped to level and the sign of the slope determines spatial position, i.e. left/right stereo channel. Thus zero crossings from pos. to neg. (neg. to pos.) become audible on the left (right) channel.

As we listen to the sound examples with the intention to uncover rhythmical patterns between the four crawl-stroke speeds, we find that there is a characteristic distribution of pitches over strokes: they begin with high pitched tones and have mainly low-pitched events at the end. This corresponds to the palm getting far away from zero-pressure early and not returning near 0 pressure for the whole time, while other arm parts experience pressure around 0 Pa, particularly the shoulders. So again, the sonification emphasizes different features than those other approaches bring into the fore.

4.6. Task-specific mapping optimizations

Finally, we present a task-specific optimized sonification that invests a bit more knowledge from the domain experts into the design. Since the pressure polarity is one of the key variables for the swimming researchers, it makes sense to represent it by a very salient parameter such as pitch. Pitch, however, is also very useful to separate and distinguish the different channels. Thus in this sonification, the sign of the pressure is responsible for a 1–2 semi-tone shift of

the 5 well separated channel-tones. The pitch values have been selected so that different combinations of polarities induces the perception of differently colored musical chords. Specifically the palm pitch (pos., neg.) was assigned to (g', a'), the tones for the hand back is (c', h), the 1/3-elbow to (g, g#), the elbow to (e, f) and the shoulder to pitch (c, H). So the hand, which is here of highest interest is assigned the highest pitch and pitch is systematically lower towards the shoulder. Sound level of these tones is an excitatory mapping from the absolute value of the derivative, and thus loud sounds indicate strong changes of pressure over time. The brightness of the timbre (i. e. bandwidth of the formant) is driven by the actual pressure value so that this information remains audible, yet appears slightly more in the background of this sonification.

Listening to the series of crawl-strokes from slow speed (sound example S6a) to sprint (S6d), we find a distinct pattern to emerge, namely that the highest pitch signal preceeds the other signals the faster the stroke becomes. Also, we become aware of harmonical patterns that correlate with the phases of the hand/arm actions. Since we cannot yet explore the sonifications in a closed interaction loop we cannot figure out what tone selections would be most suitable to turn characteristic pressure profiles for more effortless propulsion into a pleasant harmony or motif. If this should be possible, swimmers could simply be asked to attend to the motif and try to make it more harmonic. Such experiments are on our roadmap for ongoing research.

5. DISCUSSION

The paper explores the sonification of pressure data from swimming research. The presented methods contribute in different ways to understand patterns in the data, as discussed in the previous section. This section aims to look at the design and cooperation cycle from a meta level.

The different methods have been developed in the order of presentation and demonstrate various 'sonic views' on the same data. From method to method, various aspects are explored: the first approach is very generic and starts from minimal explicit knowledge; subsequent approaches invest particular domain- and task-oriented context, e.g. to turn the sonification more ergonomic for interactive use by using excitatory mappings. We found different things interesting while listening to the different sonifications, yet a lack of 'direct experience', i.e. to listen to the sonifications while swimming, makes it difficult to optimize the mappings further. So we regard these first explorations more as preparation to get a clearer feeling how to proceed once we can sonify pressure changes for the swimmer in situ. In one example, we synchronized the sonification to a video animation, and immediately felt that this makes it much easier to connect movement actions and (pressure / audible) effects.

The sonifications have not yet been optimized for aesthetics or compatibility with the soundscape of swimmers. This will become important not only for any practical use in teaching and training, but also much before, when trying to convince sportsmen and funding agencies to invest in this idea. It is, however, of lower interest if the main purpose is scientific discovery, e.g. to discover unknown relevant patterns in the data.

6. CONCLUSION

This paper contributes a new perspective on sonification as a feedback-channel for the user's action on different levels, ranging from the action level to the effect level. While the end points of this continuum have been explored in other work, we suggest the sonification of an *intermediate level* as something that we believe to be very relevant for scaffolding the learning, training and optimization of actions. For mastering or optimizing complex movements, all information levels on the continuum may be important at different stages. Thus, *multi-level sonifications* that convey information from all the levels (kinetics, intermediate effects and end effect) may be the most versatile approach, and even more so if the user or trainer can adjust the sound levels to let the most useful information stream stand out in the display as needed.

We have selected pressure data from crawl-swimming as they are an intermediate structure where we know from domain research that they matter greatly for optimizing self-propulsion. The sonifications in this paper were computed from pre-recorded data, yet the systematic variation of speed, and the availability of various executions of crawl-strokes at each swimming speed allows the listener to get an impression of what information the sonification is capable to offer. Finally, with this paper we have also documented an exploratory phase and gained some insight and gave an example how to organize research at the interface.

The next steps will be to optimize selected methods at hand of feedback from swimmers and other potential users (trainers, swimming researchers), to create sonified videos that will allow swimming researchers to better interrelate actions, data and sound, and to work towards a first real-time pressure sonification that allows us to experience the sonification while swimming. On the way we hope for discoveries and surprises.

7. ACKNOWLEDGMENT

We thank the German Research Foundation (DFG) and the Center of Excellence 277 Cognitive Interaction Technology (CITEC) that enabled this work within the German Excellence Initiative.

8. REFERENCES

[1] Schaffert, N., Mattes, K. & Effenberg A. O. (2010). Listen to the boat motion: acoustic information for elite rowers. In R. Bresin, T. Hermann, and A. Hunt, editors, Proceedings of the 3rd Interactive Sonification Workshop (ISon 2010), Stockholm, Sweden, 31–37.

[2] Effenberg, A.O. & Mechling, H. (2003). Multimodal Convergent Information Enhances Reproduction Accuracy of Sport Movements, Proc. 8th Ann. Congress of the European College of Sport Science (ECSS), ECSS, 196–197.

[3] Godbout, A., & Boyd, J. E. (2010). Corrective sonic feedback for speed skating: A case study. Proceedings of the 16th International Conference on Auditory Display (ICAD2010) Washington, DC. 23–30.

[4] Hummel J., Hermann T., Frauenberger C. & Stockman T. (2010). Interactive sonification of german wheel sports movements. Proc.ISon 2010, 3rd Interactive Sonification Workshop, KTH, Stockholm, Sweden, April 7, 2010. 17–22.

[5] Grosshauser T., Spieth, C., Bläsing, B., and Hermann, T. (2012). Wearable sensor-based real-time Sonification of motion and gesture in dance teaching and training. J. Audio Eng. Soc., (accepted for publication, to appear 2012).

[6] Grond, F., Hermann, T., Verfaille, V., and Wanderley, M. M. (2009). Methods for effective sonification of clarinetists' ancillary gestures. In Kopp, S. and Wachsmuth, I., (eds.), Gesture in Embodied Communication and Human Computer Interfaces: Proc. 8th Int. Gesture Workshop, LNCS, Berlin, Heidelberg. Springer Verlag.

[7] Höner, O., Hunt, A., Pauletto, S., Röber, N., Hermann, T., and Effenberg, A. O. (2011). Aiding movement with sonification in 'exercise, play and sport'. In Hermann, T., Hunt, A., and Neuhoff, J. G., editors, The Sonification Handbook, chapter 21, p. 525–553. Logos Publishing House, Berlin, Germany. Höner, O. (chapter ed.).

[8] Schack, T. & Mechsner, F. (2006). Representation of motor skills in human long-term memory. Neuroscience Letters, 2006, 391, 77–81.

[9] Toussaint H. M., v. d. Berg, C. & Beek, W. J. (2002). Pumped-up propulsion during front crawl swimming. Med Sci Sports Exerc. 34, 314–319.

[10] Loetz C., Reischle K. & Schmitt, G. (1988). The evaluation of highly skilled swimmers vai quanitative and qualitative Analysis. In: Swimming Science V, (Ed. by B. E. Ungerechts, K. Reischle & K. Wilke), Human Kinetics, Champaign, IL. 361–367.

[11] Takagi, H. & Wilson, B. (1999). Calculating hydrodynamic force by using pressure differences in swimming. In: K. L. Keskinen, P. V. Komi, & A. P. Hollander (Eds.), Biomechanics and Medicine in Swimming VIII. Finland: University of Jyvaskyla. 101–106.

[12] Loetz C., Reischle K. & Albrecht C. (1984). The dynamics of patterns in swimming. Leistungssport 6, 39–43.

[13] Van Manen, J. & Rijken, H. (1975). Dynamic measurement technique on swimming bodies at the Netherlands ship model basin. In: Swimming II, (Ed. by L. Lewillie & J. P. Clarys), University Park press, Baltimore. 70–79.

[14] Videler, J. J., Müller, U. K., & Stamhuis, E. J., (1999). Aquatic vertebrate locomotion: wakes from body waves. J. Exp. Biol. 202(Pt 23):3423–30.

MULTI-DIMENSIONAL SYNCHRONIZATION FOR RHYTHMIC SONIFICATION

Jeffrey E. Boyd and Andrew Godbout

Department of Computer Science
University of Calgary
Calgary, AB, Canada T2N 1N4
boyd@cpsc.ucalgary.ca, agodbout@ucalgary.ca

ABSTRACT

Human locomotion is fundamentally periodic, so when sonifying gait, it is desirable to exploit this periodicity to produce rhythmic sonification synchronized to the motion. To achieve this rhythmic sonification, some mechanism is required to synchronize an oscillator to the period of the motion. This paper presents a method to synchronize to multidimensional signals like those produced by a motion capture system. Using a subset of the joint-angle signals produced by motion capture, the method estimates the phase of a periodic, multidimensional model to match data observed from a moving subject. It does this using an optimization algorithm applied to a suitable objective function. We demonstrate the synchronization with data from a publicly available motion capture database, producing sonifications of drum beats synchronized to footfalls of subjects. The method is robust and shares some common features of phase-locked loops used for synchronizing one-dimensional sinusoidal signals. We foresee applications to sonification for athletics and clinical treatment of gait disorders.

1. INTRODUCTION

Human locomotion is, by necessity, periodic in nature [1]. Walking, jogging, running, rowing, and skating are common examples in which periodic repetition of motions move a person. We seek to use sonification to assist the training of athletes and in the clinical treatment of gait disorders. Given the periodic nature of locomotion, it then seems natural (possibly even required) to exploit this periodicity in sonification. This requires that the sonification system operate synchronously with the motion, resulting in *rhythmic sonification*.

Figure 1 illustrates the concept of rhythmic sonification. A phase signal, $\phi(t)$ (normalized such that $0 \leq \phi < 1$) provides a temporal base indicating where a subject is in the cycle of a walking stride (or other periodic motion). As $\phi(t)$ passes a phase threshold, ϕ_T, it triggers a sonic event. For example, one can select ϕ_T to correspond to the right footfall resulting in a sound that occurs synchronously with the rhythm of the walker. $\phi(t)$ is the foundation upon which one builds rhythmic sonification – once $\phi(t)$ is established, a plethora of options for rhythmic sonification becomes available.

Godbout and Boyd [2] give an example of rhythmic sonification in speed skating. They measure the ankle angle of a skater over time and synchronize to a model to generate a $\phi(t)$, and use that to provide rhythmic audio feedback to the skater. However, ankle angle measured over time is a one-dimensional signal. In contrast, motion capture systems generate many channels of data that we may wish to synchronize to. For example, the skeletal

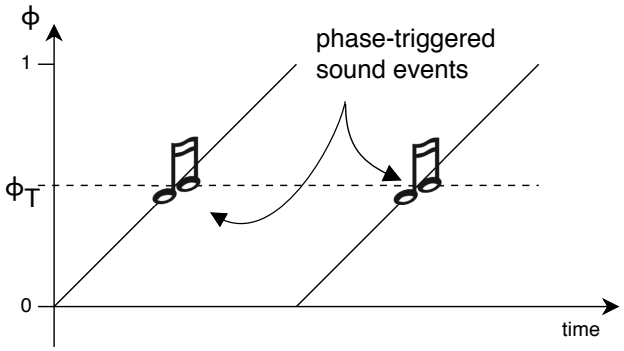


Figure 1: Phase-triggered sound events. Phase cycles from zero through one over the course of one period of the gait (or other periodic motion). As the phase passes a threshold, ϕ_T, it triggers a sound event to give rhythmic sonification synchronized to the motion.

poses measured by a Vicon [3] system in the CMU Motion Capture Database [4] contains 62 channels of data. Boyd and Sadikali [5] describe a rhythmic sonfication system using multiple channels of pixel data, but each channel is synchronized separately.

In this paper, we present a novel synchronization method to produce a *synchronized time base from multi-dimensional motion capture data*. Using multidimensional data not only provides a more reliable synchronization, but opens the doors to rhythmic sonification with numerous sensors beyond motion capture system, e.g., multi-axis accelerometers and gyros. We demonstrate our method with examples of walking and running motion capture data. The method provides a reliable time base along with a measure indicating the quality of synchronization at any point in time.

2. BACKGROUND

The synchronization of periodic events is a common phenomenon [6]. Synchronization shows up in electrical and mechanical systems, mathematics, psychology, and biological systems.

Phase-locked loops (PLL) [7] are a well known mechanism for synchronizing sinusoidal signals. PLLs are essentially feedback control systems that adjust the frequency of an internal sinusoidal oscillator to synchronize to an external oscillation. They are widely used in communications systems. Ijspeert et al. [8] and Pongas et al. [9] give examples of multi-dimensional synchronization in robotics. They measure and model periodic motions to

build control systems that allow robots to duplicate these periodic actions.

While PLLs synchronize a single oscillator, Strogatz et al. [10, 11, 12] examined the mutual synchronization of multiple oscillators. Inspired by natural phenomena such as the synchronization of fireflies, they established the regions within the space of coupling parameters that result in synchronization.

The importance of synchronization has been observed in the psychology literature. For example, Bertenthal and Pinto [13] use moving light displays to show the importance of phase locking in the perception of human gaits. When phase locking of the lights is perturbed, observers do not readily perceive a gait.

In biological systems, McGeer [1] showed that periodicity in human locomotion is an inevitable and natural consequence of the structure of the human body – gait is a limit cycle arising from body mechanics. Glass [14] examines possible mechanism for synchronization in biological structures. Cariani [15, 16, 17, 18] describes temporal coding mechanisms for perception of sound.

The message is clear – where moving people are concerned, synchronization is important. Therefore, when one seeks to sonify human motion, synchronizing to the motion is important, perhaps even necessary and we see examples in the work of Staum [19], Hamburg and Clair [20], Godbout and Boyd [2], and Boyd and Sadikali [5].

3. SYNCHRONIZATION BY OPTIMIZATION

Let $\mathbf{y}(k) = [y_1(k) \dots y_{n_c}(k)]^T$ be a vector of measurements of a periodic n_c-dimensional signal at time interval k. For example, Figure 2(a) shows an example walking gait from the CMU Motion Capture Database [4], $n_c = 4$. Note that although the full data set has 62 channels, we use only a subset for the synchronization. We choose the subset to contain those channels we expect will be best for synchronization. For example, hand and wrist movements are likely to confound the process, while McGeer [1] suggests that leg motion must be periodic. Therefore, we use the left and right femur and tibia, and take only the channels corresponding to motion in the sagital plane (x-axis rotation as denoted in the database). This corresponds to rotation about the hip and knee joints. In the remaining discussion, we assume that each channel of $\mathbf{y}$ is zero-mean, or has been preprocessed (with a high-pass filter) so that it is zero-mean. Our multidimensional synchronization process follows these steps.

1. Build a multidimensional periodic model of the motion we wish to synchronize to. This needs to be done only once for any type of motion (e.g., walking or running).
2. For an unknown signal, match the signal to the model at any point in time to estimate the phase.

The following subsections describe these steps in detail.

3.1. The Model

Let $\mathbf{y}_e(k)$ be an exemplar signal with n_s samples for the motion we wish to synchronize with. It must contain at least one full period of the motion. Our goal is to build a model function, $\mathbf{f}(\phi(k))$, that approximates $\mathbf{y}_e(k)$. Equivalently, we want n_c models such that $f_i(\phi(k)) \approx y_i(k)$ for $1 \leq i \leq n_c$.

Taking inspiration from Ijspeert et al. [8], we build f_i from a linearly weighted combination of circular Gaussian basis functions. That is:

$$f_i(\phi) = \sum_{j=1}^{n_m} w_{ij} g(\phi; \mu_j, \sigma), \tag{1}$$

where n_m is the number of Gaussian basis functions in our model, w_{ij} is weight of the j^{th} Gaussian for the i^{th} channel, and

$$g(\phi; \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(\phi-\mu)/2\sigma^2}, \tag{2}$$

is the Gaussian probability density function with center μ and standard deviation σ.

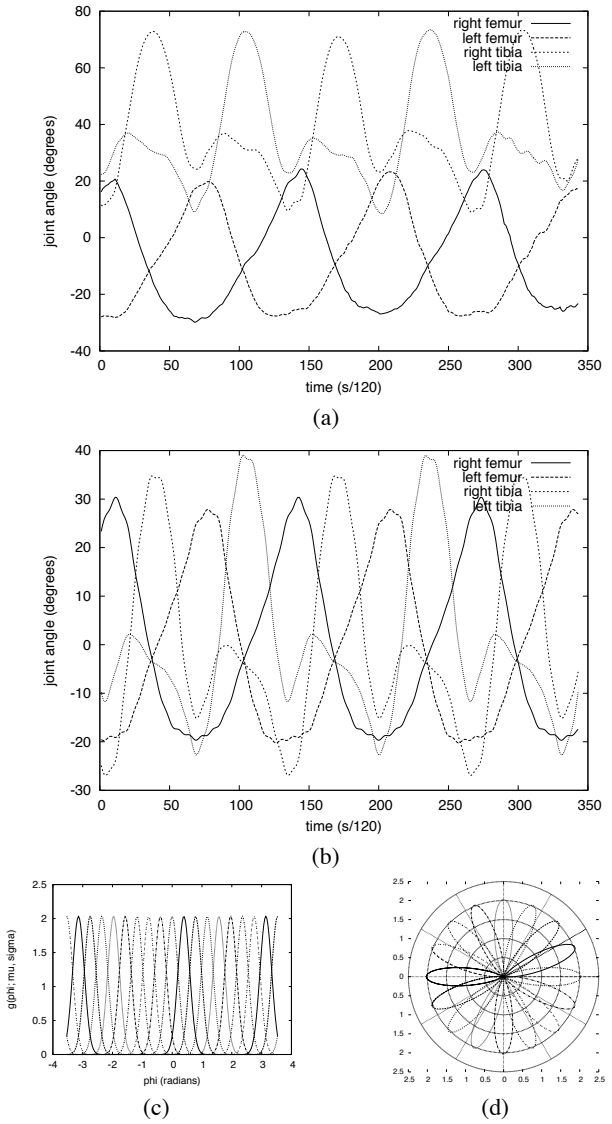


Figure 2: Signals used in the construction of a multidimensional periodic model: (a) the x-axis rotation of the left and right femur and tibia, (b) the model obtained for $n_c = 4$ and $n_m = 16$, (c) the periodic basis functions for $n_m = 16$, and (d) the same basis functions plotted in polar coordinates.

We select the μ_j such that the Gaussian bases are uniformly distributed between $-\pi$ and π, and separated by 2σ, i.e., $\sigma = \pi/n_m$. Larger values of n_m give a basis set that models $\mathbf{y}$ in more (high-frequency) detail, and lower values for n_m lead to a model of $\mathbf{y}$ that is smoother and has less high-frequency detail. Figure 2(c) and (d) show the basis functions for $n_m = 16$.

The set of w_{ij} for $1 \leq i \leq n_c$ and $1 \leq j \leq n_m$ defines our model function. To obtain the w_{ij} from $y_i(k)$, for $1 \leq k \leq n_s$, we build the following system of equations,

$$\begin{bmatrix} g(\phi(1);\mu_1,\sigma) & \dots & g(\phi(1);\mu_{n_m},\sigma) \\ \vdots & \ddots & \vdots \\ g(\phi(n_s);\mu_1,\sigma) & \dots & g(\phi(n_s);\mu_{n_m},\sigma) \end{bmatrix} \begin{bmatrix} w_{i,1} \\ \vdots \\ w_{i,n_m} \end{bmatrix} = \begin{bmatrix} y_{e_i}(1) \\ \vdots \\ y_{e_i}(n_s) \end{bmatrix}, \quad (3)$$

and solve using least squares. To get $\phi(k)$, we arbitrarily select an easily identified point in $\mathbf{y}_e$ and use that to establish a phase reference such that ϕ ramps from zero to 2π over each period of $\mathbf{y}$. For what follows, we use the first two zero crossings of the x-axis rotation of the right femur with positive slope. Figure 2(b) shows the model obtained for the exemplar in Figure 2(a) for $n_c = 4$ and $n_m = 16$.

3.2. Synchronization

To synchronize $\mathbf{f}$ with an unknown $\mathbf{y}(k)$ at sample interval k, we maximize an objective function parameterized by phase. We begin with the following:

$$E_1(\phi) = \mathbf{f}(\phi) \otimes_{\text{norm}} \mathbf{y}(k), \quad (4)$$

where $\otimes_{\text{norm}}$ denotes normalized cross-correlation.

Maximizing E_1 works well to estimate phase, but often the phase estimates deviate because the subject is not exactly like the exemplar. To *smooth out* the phase estimates, we introduce a second term to our objective function to favour solutions with a constantly increasing phase:

$$E_2(\phi) = \left(\frac{\phi - (\hat{\phi}(k-1) + \Delta\phi)}{2\pi} \right)^2, \quad (5)$$

where $\hat{\phi}(k-1)$ is the phase estimate for the previous sample of $\mathbf{y}$, and $\Delta\phi$ is the expected phase change between samples based on a typical walking cadence. Minimizing E_2 produces a phase ramp that corresponds exactly to the $\Delta\phi$. We combine E_1 and E_2 to get the following objective function,

$$E(\phi) = E_1(\phi) - \lambda E_2(\phi), \quad (6)$$

where λ is a regularization parameter. When, λ is small, the estimated phase depends primarily on a matching data to the model, and when λ is large, the estimated phase reflects only the cadence defined by $\Delta\phi$, i.e., a period of

$$\frac{2\pi T}{\Delta\phi}, \quad (7)$$

Where T is the sample period. To estimate the phase we compute

$$\hat{\phi} = \underset{\phi}{\text{argmax}}\, E(\phi), \text{and} \quad (8)$$

$$E_{\max} = \max_{\phi} E(\phi), \quad (9)$$

where $\hat{\phi}$ is our phase estimate and $E_{\max}$ is a measure of quality of match between signal and model.

As might be expected, $E(\phi)$ is periodic itself, and some care is needed to perform the optimization in the previous equation. We developed the following algorithm to compute $\hat{\phi}$.

1. Compute E on the n_m centers of the Gaussian basis functions, i.e., evaluate $E(\mu_1)$ through $E(\mu_{n_m})$.
2. Find the maximum value of E, $E_(\mu_{j_{\max}})$ among the samples in step 1.
3. Interpolate to find the position of the maximum among the samples $E(\mu_{j_{\max}-1})$, $E(\mu_{j_{\max}})$, and $E(\mu_{j_{\max}+1})$.

To interpolate between samples, we use Nishihara's [21] *subpixel interpolation* method illustrated in Figure 3. Three adjacent, uniformly spaced samples centered at the origin, $x = -1, 0, 1$, bracket a maximum of $f(x)$. The three points define a parabola. Some basic calculus reveals that the position of the maximum, x_m is at

$$x_m = \frac{-b}{2a}, \quad (10)$$

where

$$a = \frac{1}{2}(f(1) + f(-1)) - f(0), \text{and} \quad (11)$$

$$b = \frac{1}{2}(f(1) - f(-1)). \quad (12)$$

The maximum value estimated by interpolation is

$$f(x_m) = ax_m^2 + bx_m + c, \quad (13)$$

where $c = f(0)$. To find $\hat{\phi}$, and $E_{\max}$, set

$$f(-1) = E(\mu_{j_{\max}-1}) \quad (14)$$

$$f(0) = E(\mu_{j_{\max}}), \text{and} \quad (15)$$

$$f(1) = E(\mu_{j_{\max}+1}), \quad (16)$$

interpolate to find x_m and $f(x_m)$, then set

$$\hat{\phi} = \mu_{j_{\max}} + x_m \frac{\pi}{n_m}, \text{and} \quad (17)$$

$$E_{\max} = f(x_m). \quad (18)$$

4. IMPLEMENTATION AND TESTING

4.1. General

We tested our method using the CMU Motion Capture Database [4]. The database contains motion capture data for multiple subjects performing different activities over multiple trials. The motion capture data is sampled at 120*Hz*, and is available with raw video of trials, video renderings of the data, and various software tools. Of the activities available in the database, we tested on the complete selection of walking and running examples.

We implemented the method in *Octave* [22], an open-source Matlab variant, then later implemented the optimization algorithm for phase matching in Pure Data [23]. In all cases, we computed the model coefficients, w_{ij}, with Octave since this needs to be done only once, prior to any sonification.

It is necessary to manually choose the exemplar from which the model is built. In the examples here, we chose a single trial for

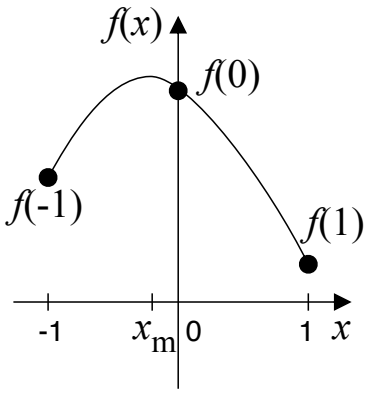
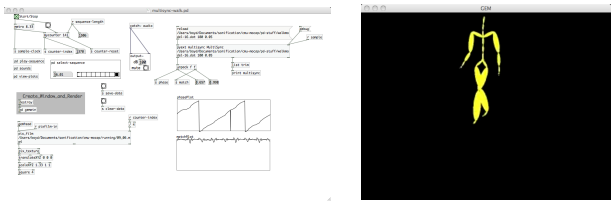


Figure 3: Nishihara sub-pixel interpolation method to find the maximum of a parabola that fits three adjacent samples with uniform spacing.



(a) (b)

Figure 4: Screenshots from Pure Data sonification patch demonstrating synchronous sonification of multidimensional data: (a) patch, and (b) synchronized video.

each of walking and running with the requirement that the exemplar sequence could contain only the activity of interest, and had to have at least two positive-going zero-crossings in the right femur x-axis rotation. The zero-crossings ensured that we could establish $\phi(k)$ correctly. While it would be possible to combine multiple subjects and trials when computing the model coefficients, it turned out not to be necessary as our results show – it seems one person's gait is similar enough to others to establish a time base. For all the examples here, walking and running alike, we used $\Delta\phi = 0.05 radians$, which corresponds to a gait period of $1.05s$. Also, for all examples, we used $\lambda = 100$.

Our sonification is simple, but sufficient to verify that we have a correct time base for other more complex sonifications. In general, once the time base is correct, timing sound events is simple. With that in mind, our sonification consists of two drum taps per gait period with the phase triggers set to correspond to the left and right footfalls. When viewing the rendered motion capture video with the sonification, it is simple to verify that the drum beats are occurring at the correct time and that the time base is correct. We normalized phases in the range $[-\pi \ldots \pi]$ to $[0 \ldots 1]$. In this case, footfalls happen at approximately $\phi_T = 0.25$ and $\phi_T = 0.75$. Figure 4 shows screenshots from the Pure Data patch in operation.

4.2. Walking

Figure 5 shows plots of $\hat{\phi}$ and $E_{\max}$ for four representative walking sequences. In all examples we tried, drum beats occurred coincidentally with footfalls in all cases where the subject was walking with a normal stride. As expected, the synchronization only fails when the subject is walking backwards or otherwise not walking normally. In these cases, the subject has deviated too far from our model gait for synchronization to occur.

Figure 5(a) shows plots for our walk training subject, i.e., it shows the model synchronizing to itself – a strawman test. The second term of Equation 6 starts with an arbitrary $\hat{\phi}$ and takes a few samples to converge to the correct phase. After this convergence, the phase is synchronized correctly. The longest convergence period we observed was approximately 75% of a gait cycle, and most often the convergence occurs in half a cycle or less. Note that the values of $E_{\max}$ are low during the convergence interval. So although the system has not converged, it has a numerical indicator that the phase estimate is not good. This example also has a period of $1.1s$, which happens to correspond closely to the natural period for $\Delta\phi = 0.05 radians$.

Figure 5(b) shows results for a similar trial, but with a different subject. This subject has a much slower stride, with a period of $1.6s$. Although this is significantly different than the natural period for $\Delta\phi = 0.05 radians$, the system correctly locks to the phase of the walker while the second term of Equation 6 smooths the phase estimates.

Figure 5(c) corresponds to a sequence in which the subject walks for a few paces, stops, turns around, and walks a few paces back to their starting position. The synchronization plots clearly show this. In the middle of the plot, there is an interval during which the the phase stops ramping and $E_{\max}$ drops which corresponds to the moment when the subject stops and turns. The sonification produces correct footfalls during the normal paces, and a couple of spurious taps as the subject stops and turns.

Walking backwards confounds the synchronization and sonification as shown in Figure 5(d). These plots correspond to part of a sequence where the subject walks backwards for a couple of paces. Clearly the synchronization has failed. The second term of the objective function (Equation 6) drives $\hat{\phi}$ forward in an approximate phase ramp, but waveform is irregular and $E_{\max}$ values are sporadically low indicating a poor match. We did try synchronizing to this sequence with the second term of Equation 6 removed, i.e., $\lambda = 0$. In this case we do see a downward phase ramp as one might expect, but the cost is in a noisier phase estimate throughout the entire sequence.

4.3. Running

Figure 6 shows plots of $\hat{\phi}$ and $E_{\max}$ for four representative running sequences. As was the case with the walking examples, the drum beats occurred simultaneously with footfalls during normal running. Most of the running sequences are by necessity shorter – the higher speed means the subject is in the field of view of the motion capture system for a shorter period of time, unless they alter their gait to change direction.

Figure 6(a) shows synchronization with the same subject used for our running model, but for a different trial. Synchronization is comparable to what we observed for walking. Figure 6(b) shows a sequence for a different subject, again exhibiting excellent synchronization. It is worth noting that although the stride frequencies for these are significantly faster than the natural frequency for $\Delta\phi = 0.05 radians$ (periods of $0.68s$ and $0.78s$ verus $1.05s$), our system still synchronizes well.

Figure 6(c) corresponds to a sequence in which the subject

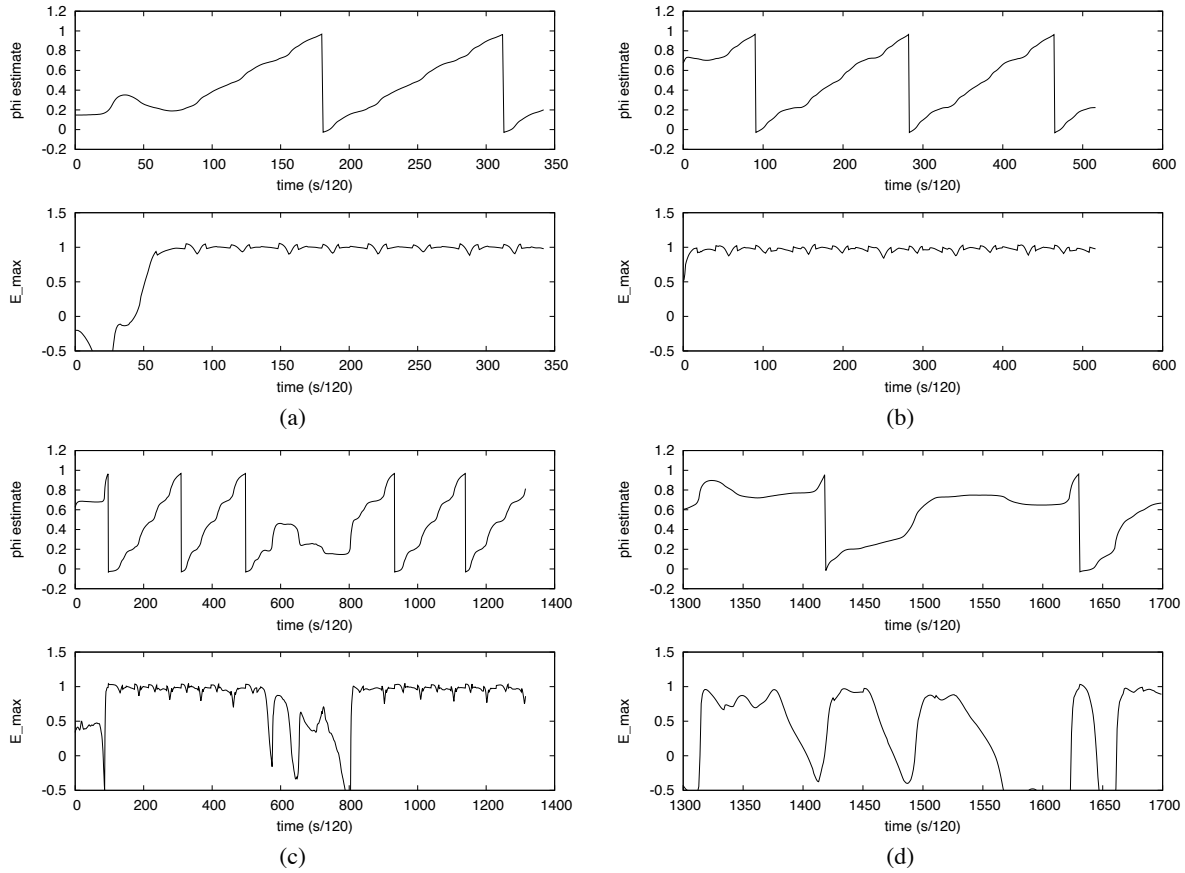


Figure 5: Walking synchronization results: (a) model squence, (b) a typical walking sequence, (c) subject stopping and turning, and (d) subject walking backward. In all examples, the upper plot shows the phase estimate, $\hat{\phi}$, and the lower plot shows $E_{\max}$.

runs, comes to a stop, and with a hop changes direction. One can see when the hop occurs where the phase ramp is distorted near the middle of the sequence, and where the sporadic drops in $E_{\max}$ occur. Again, it was in these sorts of variations from a normal running gait where the sonification of footfalls becomes erratic.

Figure 6(d) shows plots for a longer running sequence in which the subject runs around the field of view in a box pattern, turning at the corners. The effects of this pattern are clear in the plots. One can see the dips in $E_{\max}$ at the corners, and also some distortion in the phase ramps as the subject alters the gait to accommodate the corner.

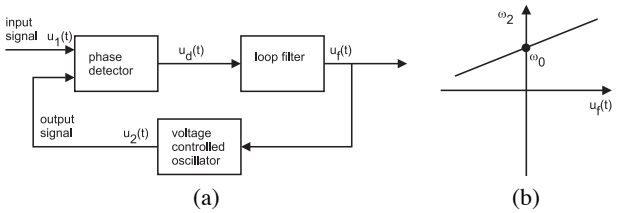


Figure 7: A basic phase-locked loop: (a) block diagram, and (b) the transfer function of the voltage controlled oscillator.

5. DISCUSSION

As a way to understand the synchronization method presented here, we can compare to PLLs. Figure 7 shows the elements of a PLL [7]. The phase comparator and the (low-pass) loop filter together compare input oscillations to the oscillations of an internal oscillator, the *voltage controlled oscillator* (VCO). The transfer function of the VCO, shown in Figure 7(b) relates the frequency of the internal oscillator to its natural frequency, ω_0, and the difference between internal and external signals. It is not meaningful to compare two one-dimensional signals instantaneously, leading to the requirement to have a low-pass filter that effectively integrates phase comparisons over time.

In their synchronization system for speed skating, Godbout and Boyd [2] also integrate a comparison over time when they compute the normalized cross-correlation over a window of one period. They have no equivalent to the VCO, relying instead on a brute-force search over frequency space for every sample.

In the system presented here, we are getting close to a multidimensional PLL for arbitrary wave forms. The E_1 term in Equation 6 compares an incoming multidimensional signal to the internal multidimensional oscillator in our model. The need for the low-pass filter is obviated by the multidimensional signal – we integrate over dimensions instead. This allows us to get an instan-

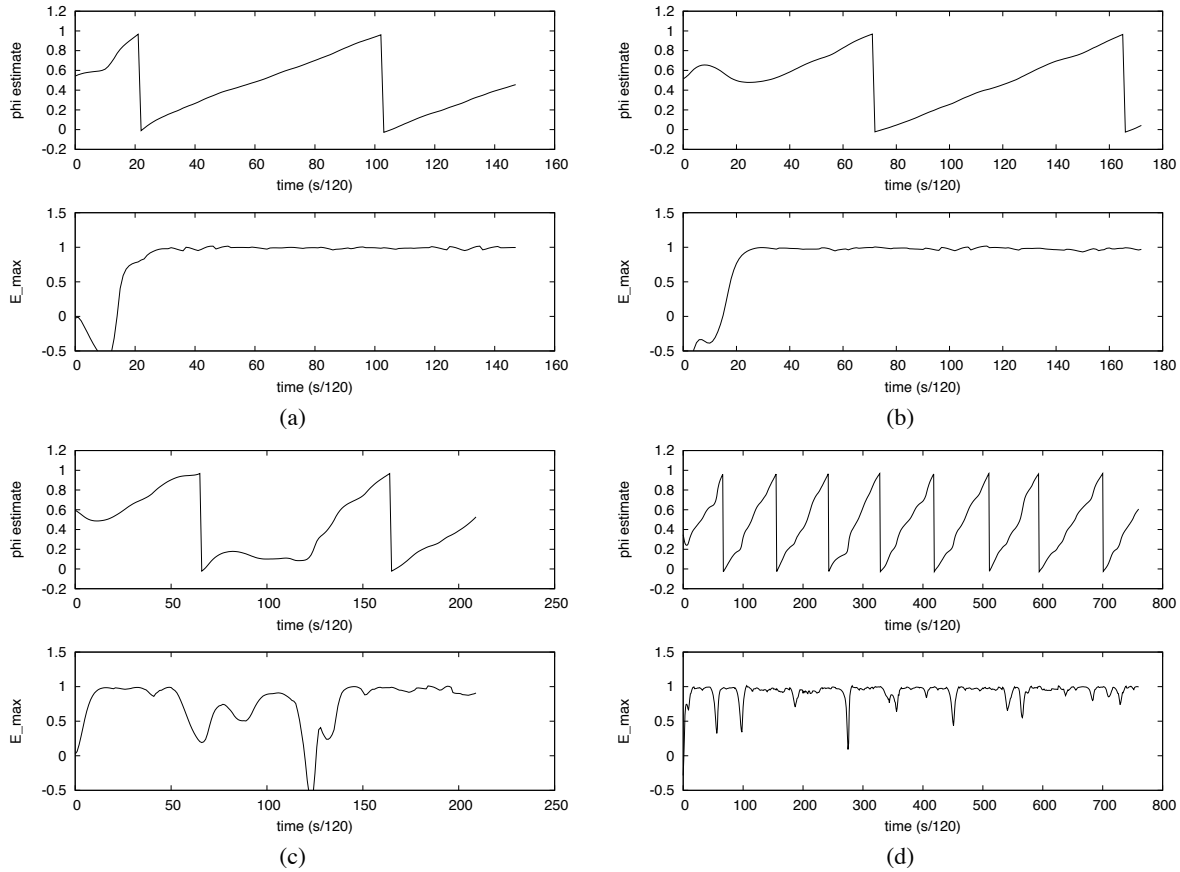


Figure 6: Running synchronization results: (a) the model sequence, (b) a typical running sequence, (c) hop and change of direction, and (d) running around a square. In all examples, the upper plot shows the phase estimate, $\hat{\phi}$, and the lower plot shows $E_{\max}$.

taneous comparison that is not possible with one-dimensional signals. Further more, the E_2 term in Equation 6 is equivalent to the VCO. It ramps at a natural frequency defined by $\Delta\phi$ but responds to the external signal when combined with E_1. Our system is not precisely equivalent to a PLL though – it lacks feedback to track the incoming signal, relying on an optimization for each sample interval.

It is important to note that although we synchronize with just four channels of the motion capture data, once we are synchronized, we can rhythmically sonify any and all channels of the data. We see potential here because:

- our method opens the door to real-time rhythmic sonification for athletics and clinical applications, and
- motion capture is getting cheaper (consider the MicroSoft Kinect) which will lower the cost requirements for using this type of sonification.

6. CONCLUSIONS

We have presented a method of synchronization applicable to periodic, multidimensional signals like those produced by motion capture systems acquiring data from locomotion. The system features key elements of PLLs, an established method for synchronizing internal oscillators to incoming sinusoids. Once this synchronization is established, it provides the temporal basis for rhythmic sonification.

7. REFERENCES

[1] T. McGeer, "Passive walking with knees," in *IEEE International conference on Robotics and Automation*, 1990, pp. 1640–1645.

[2] A. Godbout and J. E. Boyd, "Corrective sonic feedback for speed skating: a case study," in *International Conference on Auditory Display*, Washington, DC, June 2010, pp. 23–30.

[3] "Motion capture systems from vicon," Retrieved January 23, 2012, from http://www.vicon.com.

[4] "Cmu graphics lab motion capture database," http://mocap.cs.cmu.edu/, created with funding from NSF EIA-0196217.

[5] J. E. Boyd and A. Sadikali, "Rhythmic gait signatures from video without motion capture," in *International Conference on Auditory Display*, Washington, DC, June 2010, pp. 187–191.

[6] S. Strogatz, *Sync: The Emerging Science of Spontaneous Order*. New York: Theia Books, 2003.

[7] R. E. Best, *Phase-locked loops design, simulation and applications*. New York: McGraw–Hill, 1999.

[8] A. J. Ijspeert, J. Nakanishi, and S.Schaal, "Learning rhythmic movements by demonstration using nonlinear oscillators," in *International Conference on Intelligent Robots and Systems*, Lausanne, Switzerland, October 2002, pp. 958–963.

[9] D. Pongas, A. Billard, and S. Schaal, "Rapid synchronization and accurate phase-locking of rhythmic motor primitives," in *International Conference on Intelligent Robots and Systems*, Edmonton, AB, Canada, August 2005, pp. 2911–2916.

[10] P. C. Matthews and S. H. Strogatz, "Phase diagram for the collective behavior of limit-cycle oscillators," *Physical Review Letters*, vol. 65, no. 14, pp. 1701–1704, October 1990.

[11] R. E. Mirollo and S. H. Strogatz, "Synchronization of pulse-coupled biological oscillators," *SIAM Journal of Applied Mathematics*, vol. 50, no. 6, pp. 1645–1662, November 1990.

[12] S. H. Strogatz, R. E. Mirollo, and P. C. Matthews, "Coupled nonlinear oscillators below the synchronization threshold: relaxation be generalized landau damping," *Physical Review Letters*, vol. 68, no. 18, pp. 2730–2733, May 1992.

[13] B. I. Bertenthal and J. Pinto, "Complementary processes in the perception and production of human movements," in *A Dynamic Systems Approach to Development: Applications*, L. B. Smith and E. Thelen, Eds. Cambridge, MA: MIT Press, 1993, pp. 209–239.

[14] L. Glass, *Nonlinear Dynamics in Physiology and Medicine*, ser. Interdisciplinary Applied Mathematics. Springer, 2003, ch. Resetting and entraining biological rhythms, pp. 123–148.

[15] P. Cariani, "Temporal coding of periodicity pitch in the auditory system: an overview," *Neural Plasticity*, vol. 6, no. 4, pp. 147–172, 1999.

[16] ——, "Temporal coding of sensory information in the brain," *Acoustical Science and Technology*, vol. 22, no. 2, pp. 77–84, 2001.

[17] ——, "Temporal codes, timing nets, and music perception," *Journal of New Music Research*, vol. 30, no. 2, pp. 107–135, 2002.

[18] ——, "Temporal codes and computations for sensory representation and scene analysis," *IEEE Transactions on Neural Networks*, vol. 15, no. 5, pp. 1100–1111, 2004.

[19] M. J. Staum, "Music and rhythmic stimuli in the rehabilitation of gait disorders," *Journal of music therapy*, vol. 20, no. 2, pp. 69–87, 1983.

[20] J. Hamburg and A. A. Clair, "The effects of a movement with music program on measures of balance and gait speed in healthy older adults," *Journal of Music Therapy*, pp. 212–226, 2003.

[21] H. K. Nishihara, "Prism: A practical real-time imaging stereo matcher," MIT AI Lab, Tech. Rep. AI Memo 780, 1984.

[22] "Octave," Retrieved February 8, 2012, from http://www.gnu.org/software/octave/.

[23] "Pure data," Retrieved February 8, 2012, from http://puredata.info/.

INTUITIVE AND INTERACTIVE MOVEMENT SONIFICATION ON A RISC / DSP PLATFORM

Hans-Peter Brückner, Matthis Wielage and Holger Blume

Leibniz Universität Hannover,
Institute of Microelectronic Systems,
Appelstraße 4, 30167 Hannover, Germany
{brueckner, wielage, blume}@ims.uni-hannover.de

ABSTRACT

A major requirement for effective and interactive sonification in rehabilitation is the availability of a mobile platform. Portable state of the art motion capturing is achieved with inertial sensors. This paper presents a real-time, low latency sonification demonstrator based on an low power consumption ARM Cortex A8 processor, which is designed for mobile usage. The sonification demonstrator is based on the Texas Instruments C6A816x / AM389x development board. It enables research in continuous real time sonification of human motion to improve the process of motion learning in stroke rehabilitation. Profiling results are used to benchmark the Integra software application against a PC based version in terms of signal processing latency. Furthermore, a new sonification mapping, basing on the beat effect, is introduced. This mapping is especially usable for people suffering from partial deafness. A subjective test series shows the understandability of this mapping for healthy subjects, in comparison to a previously proposed sonification mapping.

1. INTRODUCTION

Several studies in the field of sports science claim that motion learning benefits from movement sonification [1]. Sonification is the displaying of non-speech information through audio signals [2]. In the rehabilitation context, benefits from interactive movement sonification have been shown [4]. Also efficacy in stroke rehabilitation is proved [4].

The proposed demonstrator is designed for usage in stroke rehabilitation. This kind of rehabilitation focuses on regaining a maximum level of independence within daily activity. Therefore, many rehabilitation exercises focus on upper extremities movements, as these are required in basic tasks, like eating, drinking and tooth brushing. Inertial sensor system set up is chosen according to [5], with one sensor at upper arm and one sensor attached to forearm. Sonification acoustically displays the wrist position, captured by inertial sensors. This provides information about movement performance.

Using movement sonification in sports or rehabilitation requires fully mobile and portable sonification systems. Depending on the chosen mapping parameters, sample based sound synthesis gets quite computational intensive. Therefore, power demanding processors are required. PC based hardware platforms [6], [7] require a high power budget and are limited to stationary usage.

For this reason an approach for real time sonification of complex movements captured by inertial sensors on a low power consumption processor platform is presented in this paper. The sonification demonstrator consists of a Texas Instruments (TI) C6-Integra processor integrated in the C6a816x/AM389x evaluation module comprising an ARM Cortex A8 processor and a Digital Signal Processor (DSP) [8]. Movements are captured with an Xsens inertial sensor system [9] consisting of MTx sensors and an Xbus Master device. The number of MTx sensors can be scaled flexible to up to ten sensors according to motion capturing demands. Speakers or headphones can be used to listen to the generated stereo audio signal. Hardware demonstrator components and structure are shown in Figure 1. Sensor data acquisition, sonification parameter calculation and audio synthesis are handled on the Cortex A8 CPU. A setup is chosen, where sonification displays the wrist position in relation to the patient's body based on different parameter mappings.

Sample based sonification is achieved using the Sound Synthesis Toolkit (STK) [10]. The STK consists of audio signal processing and synthesis classes in C++. Thus, it allows seamless integration in the C++ based sensor system application programming interface (API) and orientation data processing framework. Different basic STK sound generators are used for sonification. The mappings are benchmarked in terms of computational latency and intuitive understandability of the sonification.

The paper is organized as follows: Section 2 presents related work. Section 3 introduces the evaluation board and the ARM processor. The proposed software architecture is explained in Section 4. Section 5 introduces the new beat effect based mapping. In Section 6, the intuitive usability of sonification mappings is evaluated. Profiling results and a benchmark against a PC based platform are given in section 7. Conclusions are given in Section 9.

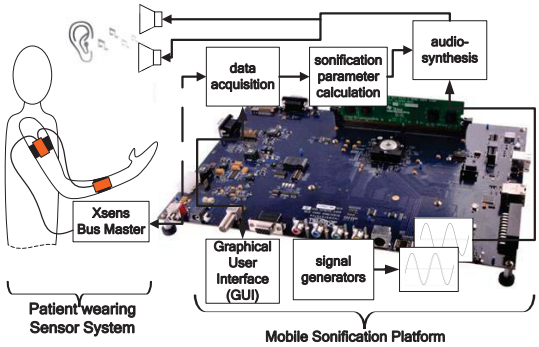


Figure 1: Hardware demonstrator structure

2. RELATED WORK

Movement sonification is explored in multiple research projects. Particularly, sonification on mobile devices is a research focus for several years. However, the proposed hardware platforms suffer from drastic limitations in capturing of complex movements and sonification design. Although, there are a variety of applications, like stroke rehabilitation, where mobile sonification of complex movements, provided by this proposed hardware platform is mandatory.

A framework designed for continuous real time movement sonification is presented in [6]. User movements are captured using an optical infrared marker based capturing system. Therefore, absolute position information is additionally provided to relative orientation information. Fully customizable sonification is achieved using Supercollider [11]. This system is not prepared for a mobile usage, because it is based on an optical motion capturing system and a desktop computer based processing.

In [12] a system for sonification of biofeedback signals is presented. Biofeedback sonification should here for example provide information to users' stress level or drowsiness. The system is capable of multiple signal sonifications. Mobile usability is achieved by operating on a Nokia N900 Smartphone with wireless connected sensors. In contrast to the work presented here, the sonification bases on basic alert signals. Additionally, there is not any complex data processing reported.

The work described in [13] generates a sonification based on captured input gestures on a PocketPC. Gestures are captured using an attached external gyroscope. The captured data is processed to identify distinct gestures and give an auditory feedback. Sonification is achieved by linking recognized gestures to very basic audio sources. Compared to the desired application proposed in this paper, this approach is not able to accurately detect and track whole arm movements and giving a complex auditory feedback.

A mobile system for improving running mechanics is developed in [14]. The system comprises a mobile phone and triaxial accelerometers and gyroscopes connected via Bluetooth. During usage, the sensor is attached to the sacrum and accelerometer data is captured. In processing steps, the runner's average center of mass is computed. Providing this information to the runner gives an objective feedback to his running technique. Due to limited computing capabilities of the chosen hardware platform, sonification is based on playback of prerecorded sound files.

Mobile sonification of sculler movements in [15] is realized using a Symbian OS [16] mobile phone. To provide information about boat velocity, a built in GPS receiver and an external acceleration sensor are used. Feedback is given via MIDI sounds. The authors report that the current approach is suffering from noticeable drift caused by accelerometer bias. In contrast to the work presented here, there is no capturing of complex, multi segment movements.

Expressive music performances are used for sonification in [17]. This work also is based on a mobile phone as hardware platform. User movements are captured via the built in accelerometer. A computation step classifies several gestures based on accelerometer data. For usage in rehabilitation context, this approach is limited, as the usage of one accelerometer only provides sparse information, when performing complex movements.

Focusing on non mobile application of sonification in rehabilitation there are numerous research activities [18], [19].

In contrast to the low latency approach proposed in this paper, in none of the platforms listed in related work, latency is considered. Overall hardware and software latency design goal is 30 ms, as higher values result in recognizable differences in visual and audio cognition [19].

3. MOBILE HARDWARE PLATFORM

The C6-Integra processor consists of an ARM Cortex A8 processor and a C674x fixed and floating point DSP, both operating at 1 GHz. As both processors and additional modules are integrated on a single die, this is called a 'System-on-Chip' (SoC). The Cortex A8 core is a Reduced Instruction Set Computer (RISC) especially designed for usage in mobile devices [21]. Reduced instruction set allows designing area and power consumption efficient processors, as there is less effort for instruction decoding required.

The Cortex A8 can achieve additional speedup by using the Single Instruction Multiple Data (SIMD) unit NEON [21]. This unit allows the computation of 16 64- and 128Bit-SIMD-instructions in parallel. It is designed for usage in audio and video processing applications to overcome the needs for custom hardware accelerators and therefore keep flexibility for future standards or different workloads. The unit is especially designed for floating point multiplications, shift and multiply accumulate operations.

Figure 2 shows a block diagram of the Integra SoC with additionally available accelerators and memory. Both processor cores communicate using a packet based communication protocol.

The TI C6A816x evaluation module (EVM) allows the connection of external devices using several interfaces, like USB and serial ports, video and audio interfaces and an SD-card slot. Due to the lack of an appropriate driver, the XBus Kit is connected via a Blueserial [19] Bluetooth to serial converter. User-friendly operation is achieved via an external 8" touch screen, connected by a HDMI cable. Linux is chosen as operating system to support audio and video drivers and the Qt [23] based application. The onboard stereo audio converter TVL320AIC3106 [20] allows direct connection to speakers or headphones. Additional available interfaces are Ethernet, SCART, S-Video, VG, IR and JTAG for debugging.

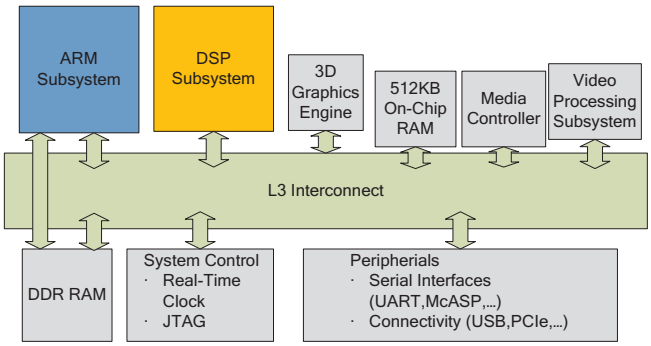


Figure 2: C6A816x System-on-Chip block diagram

4. SOFTWARE ARCHITECTURE

The proposed software provides auditory feedback of the wrist position in three-dimensional space. Detecting movements with up to ten inertial sensors and other processing steps enables the sonification of a variety of motion parameters, like segment accelerations, velocities, angles and relative positions. In addition, there is a graphical user interface (GUI), which visualizes movement features and allows control of the sonification process and parameters. For example, different mappings from parameter to sound can be chosen here.

The interactive human movement sonification software is based on the object oriented programming language C++. The GUI is based on the C++ class library Qt [23], which extends C++ to skills for GUI design and inter-object communication.

The application is characterized by a multi-threaded architecture. Thus, basic tasks are logically separated and run in multiple threads, basing on the producer-consumer concept.

In terms of the sonification application, the producer thread communicates with the Xsens hardware. The data of the inertial sensors is requested and then stored in a shared memory. The consumer thread retrieves the data, removes it from the queue and starts processing. The advantage of this design pattern is that the processing of data does not block the whole system, and also allows limited parallelism. The producer can obtain the data, while the consumer is running working tasks. Furthermore, an adaptation of different clock speeds is possible. For example, the intertial sensor data rate is 100 Hz, while audio samples are generated at 44.1 kHz. This allows a higher throughput, which is required for a low latency, real-time implementation of the demonstration software [5].

Figure 3 shows the class structure within the software architecture in a Block diagram.

The XsensData class represents the producer thread and communicates with the sensors on the XsensCMT library. The library handles low-level communication with the sensors. Received sensor data packets are written to a queue and the HandleData class (consumer thread) performs the processing. The wrist position vector is generated from a weighted normalized vector addition of the individual arm segments.

Coordinates system and sensor positions are chosen according to [5]. Cartesian coordinates and radius are normalized to the test subjects arm length.

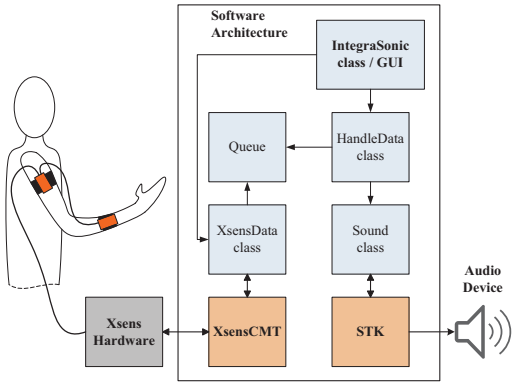


Figure 3: Software architecture

The data is then passed to the class Sound for sonification. The class Sound handles the control and the generation of the audio stream using the Sound Synthesis Toolkit (STK). The initialization of the GUI and the initializing of slots and signals are performed by the class IntegraSonic. Furthermore, this class controls the threads, as it is the main class.

4.1. Software optimization steps

Due to the lower operation frequency of the Integra processors Cortex A8 processor core, in contrast to the development PC, software optimization was performed to keep the overall latency constant. Therefore, functions with large processing times and most frequent calls, identified by software profiling, were optimized.

Since the queue was identified to have major impact on the processor load, two approaches were implemented to reduce this burden. First, the class QQueue of the Qt framework has been replaced; second the extension QWaitCondition has been integrated, to stop trying to poll data items when the Queue is empty. The originally used class QQueue was replaced by a simplified queue class, which contains only the most basic functions. These are:

- Adding an element to the queue
- Removing an element from the queue
- Check that objects are present in the queue.
- Number of elements in the queue

The items in the queue are inserted as objects of class QueueElement, which include not only the item itself, but also have a pointer to the next element.

QWaitCondition (an extension of the Qt framework) was integrated into the application to allow a better synchronization of threads, to reduce computational load.

This extension allows threads to signal another thread that a certain condition is met. Thus, an instruction can hold a thread until another thread calls a wake.

Within the application this functionality is carried out by the XsensData class; when data is stored in the queue, it wakes the HandleData thread by calling the function wakeAll(). The HandleData thread can now remove the data from the queue and performs computation. When the thread task is finished, a sleep state is obtained by calling the function wait(). This sleep state again is terminated when waking is performed, or 10ms have passed. (10 ms = sensor system sampling interval)

5. PARAMTER MAPPING

Presenting movement information for stroke patients via sonification has to ensure being understandable and intuitive for these persons. Therefore, mappings using stereo effects might be impractical, as [25] shows a large impairment in audio perception of stroke patients. The study reported that significant problems in stroke patients passing the dichotic competing sentence testing (DCST) occurred. Therefore, the stereo effect based mapping presented in [5] does not fit to the requirements. The new proposed beat effect mapping considers these effects, therefore it is limited to frequency and volume based sonification mappings.

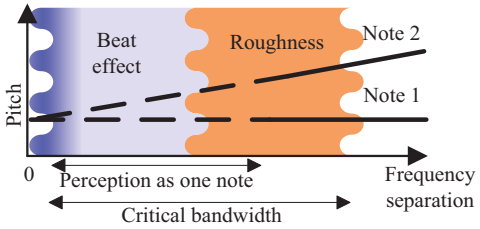


Figure 4: Acoustical beat effect in relation to frequency separation [26]

Different sound frequencies are assigned to relatively broad excitation zones in the ear, so that in case of low frequency differences, the corresponding excitation zones overlap. Thereby, the psychoacoustic beat effect is generated. Figure 4 shows the human perception influenced by frequency separation. Pitch indicates the sine generators base frequencies.

Acoustical beat is realized using two sound synthesis toolkit [10] sine generators, operating at slightly different frequencies according to [26]. As this kind of sonification does not rely on stereo effects for displaying information, it is also applicable in rehabilitation of stroke patients with partial deafness. The general concept is shown in Figure 5.

For frequency differences up to 10 Hz, the tones are perceived as volume fluctuations, corresponding to the mean of the frequencies. Further increases result in a perception of quick succession of beats, which blend at above 15-20 Hz difference to one tone at a constant volume with a rough sound character. This roughness increases up to a frequency deviation of 10% and then falls, until two harsh sounds are perceived. Exceeding the critical bandwidth this roughness disappears. The critical bandwidth is in the range of a major and a minor third.

For both coordinate systems, the origin is located at shoulder joint and wrist position is computed assuming a rigid body [5]. A test series is set up to show if coordinate system choice influences intuitive understandability of the sonification mapping. Finally, conclusions are given by comparing the proposed beat effect sonification against a sonification based on a single sine generator and an artificial instrument in terms of computation effort, intuitive understandability and ambience.

Instrument	Volume	Base frequency
Cartesian (x, y, z)	amplitude (A) = 0.8-0.5 * y left channel volume = A * (⅔ * x + ⅓) right channel volume = A * (- ⅔ * x + ⅓)	ranging from a (z < -0.92) to as'' (z > 0.92) in steps of 0.09 on a chromatic scale (a=220 Hz; as''=830.6 Hz)
Spherical (r, φ, θ)	amplitude (A) = 2-1.8 * r left channel volume = A * (φ – ⅓ π) right channel volume = A * (φ – ⅔ π)	ranging from a (θ > 2.42 rad) to as'' (θ < 0.79 rad) in steps of 4° on a chromatic scale (a=220 Hz; as''=830.6 Hz)

Table 1: Instrument sonification parameters

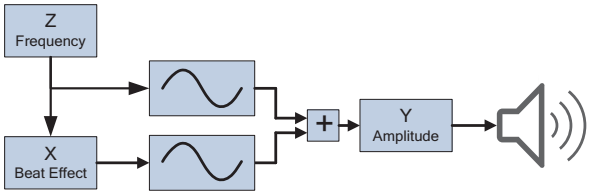


Figure 5: Beat effect realization

Beat effect	Volume	Base frequency	Frequency difference
Cartesian (x, y, z)	volume = 0.3+0.7*abs(y)	frequency= z*3300+550	diff= (x+1)*10
Spherical (r, φ, θ)	volume = 0.3+0.7*abs(θ/π)	frequency= r*330/π+550	diff= φ *10/π+10

Table 2: Beat effect sonification parameters

6. EVALUATION OF THE INTUITVE UNDERSTANDING OF SONIFICATION

A subjective test series with 40 participants was set up to compare sonification mappings according to [5] (Instrument based wrist position sonification based on spherical coordinate system, later referred as A) and the proposed beat effect mapping. Furthermore wrist position information was provided using a Cartesian and a spherical coordinate system. Participants were encouraged to report if they were able to identify movement influence on the generated audio signal and rate the acceptability (pleasant and encouraging sound). Therefore, participants were blindfolded to constrain movement perception to auditory and proprioceptive information.

6.1. Subjects

The subjects participating in the study were 36 male subjects and 4 female subjects between 16 and 31 years. Only non experts were questioned. To suppress learning effects, the presented mapping order was randomized. Persons with previous experience in movement sonification were identified. The questionnaire was designed according to ITU-R recommendations for subjective sound quality assessment [27]. In order to achieve a good sound quality, Sennheiser PXC310 headphones were used in a configuration according to Figure 6.

6.2. Test Setup

Sonification setups according to Table 3 were presented in a randomized order to the subjects. During 45 seconds, the participants were asked to perform free movements and try to discover to influence of movements within the sonification mapping without any previous knowledge.

Identifier	**Sonification Mapping**	**Coordinate System**
A	Instrument	Spherical
B	Beat effect	Spherical
C	Instrument	Cartesian
D	Beat effect	Cartesian

Table 3: Evaluated sonification mapping setups

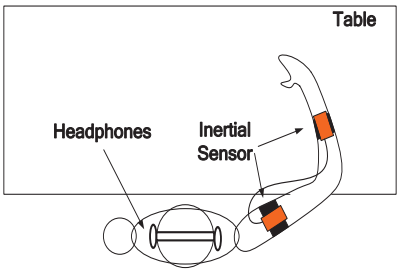


Figure 6: Setup used for evaluation

The questionnaires datasets were submitted to two-way analyses of variances (ANOVA) with the between-factor Group and the within-factor treatment. Post hoc comparisons were made with Fisher's LSD-tests. Independent one-sample t-test was used to identify significant differences of the mean values in the understandability evaluation in comparison to the "No Correlation" statement.

6.3. Questionnaires Design

Test subjects were asked to rate the acceptance and understandability of the four different parameters to sound mappings. Acceptability had to be rated on a four point scale ranging from comfortable to annoying. The understandability of the presented movement information was rated on a five point scale ranging from clearly perceptible to no correlation.

After performing each of the four test trials, the test subjects answered the questions according to acceptance and understandability. Finally, the test subjects were asked to chose their favorite mapping according to understandability.

Table 4 and Table 5 give the interpretation of results shown in further figures and the questionnaires ratings.

Rate	**Coding**
Comfortable	1
	2
	3
Annoying	4

Table 4: Acceptability (comfort) evaluation mapping

Rate	**Coding**
Clearly Perceptible	1
Perceptible	2
Moderate Perceptible	3
Hardly Perceptible	4
No Correlation	5

Table 5: Understandability evaluation mapping

6.4. Subjective Test Series Analysis

Results of the survey after questioning 40 subjects are given in Table 6. Evaluation shows that Sonification C (Instrument; Cartesian coordinates) was rated as the most pleasant mapping. Regarding understandability, test subjects rated Sonification A (Instrument; Spherical coordinates) best.

In coincidence with the observations in Figure 7, ANOVA of the acceptability evaluation showed a significant effect of the different sonification mapping A-D ($F_{(3,117)}$=8.92, $p < 0.001$ η^2=0.314). Post hoc analysis of the acceptability evaluation confirmed, that mapping A significantly differs from B ($p < 0.05$), and B significantly differs from all others ($p < 0.05$), and C significantly differs from B and D ($p < 0.05$), and D significantly differs from B and C ($p < 0.05$).

In accordance with the observations in Figure 8, ANOVA of the understandability yielded a significant effect of the sonification mapping ($F_{(3,117)}$=13.30, $p < 0.001$, η^2=0.462). Post hoc analysis of the understandability evaluation confirmed that sonification mapping A significantly differs from B and C ($p < 0.05$), and B is significantly different from all others ($p < 0.05$) and C significantly differs from A and B ($p < 0.05$), and also D significantly differs from B ($p < 0.05$).

Students t-test confirmed, that all sonification mappings differ significantly from 5 ("No Correlation"), (A: $t_{(39)}$=-24.60,.B: $t_{(39)}$=-15,77, C: $t_{(39)}$=-23.80, D: $t_{(39)}$=-22.80, with $p < 0.001$).

Identifier	Acceptability		Understandability	
	mean	sd	mean	sd
Sonification A	2.00	0.78	1.63	0.87
Sonification B	2.63	1.00	2.63	0.95
Sonification C	1.88	0.82	1.83	0.84
Sonification D	2.25	0.93	1.95	0.85

Table 6: Survey results

Figure 7 shows results of the acceptability evaluation with the corresponding error bars of the sonification according to Table 3. The results show that most test subjects favor the instrument and stereo effect based mappings A and C. Only one test subject could not find any correlation while performing the free trial using these mappings. All others found the mappings to be at least moderate perceptible.

Analysis of the understandability evaluation of the sonification according to Table 3 in Figure 8 shows, that also here the artificial bowed instrument based sonification was rated best. The beat effect based sonification shows remarkably results when using a Cartesian coordinate system. In contrast to beat effect based sonification, in instrument based sonification there is only a small difference in understandability, dependent on the coordinate system. The beat effect showed significantly better results when using a Cartesian coordinate system for wrist position calculation.

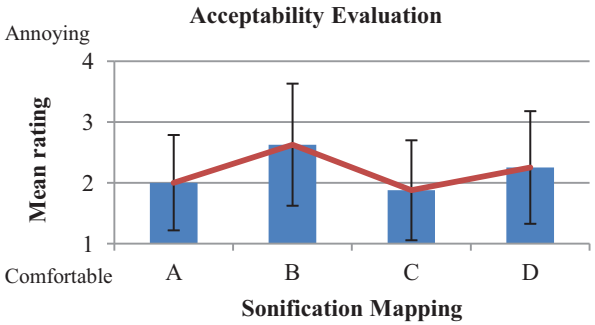


Figure 7: Acceptability evaluation

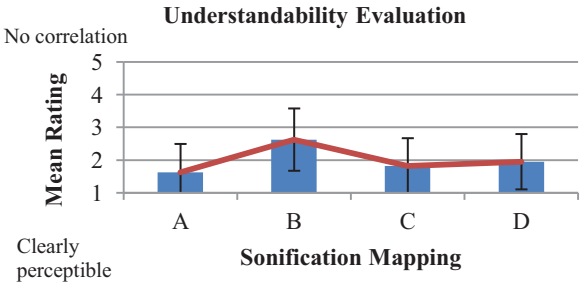


Figure 8: Rating of the individual sonification mappings

After finishing all four free trials the test subjects were asked to vote for their most favorite sonification. Figure 9 shows the rates of this survey. It becomes clear, that instrument sonification is preferred by unimpaired subjects.

7. SOFTWARE BENCHMARK

According to [5] the system latency is divided into three blocks. The first fraction is the data acquisition time of the sensors and the transmission time from the sensor system to the host platform, here there is less possibility for latency minimizations as it is limited by the Xsens sensor system itself. Latency induced by computations on the hardware platform, as PC or TI Integra, is represented by the second part. Finally, the last part consists of delay caused by the minimum required audio buffer size, either by using Microsoft DirectSound or Linux ALSA.

For profiling under Linux gprof was used. This profiler only allows sampling based profiling, which means that the processors call stack is evaluated at distinct sampling intervals. To provide accurate information using this statistical profiling method, a log-file of 28,882 samples was used.

The benchmarked development PC, used for reference value generation, is equipped with an Intel Core2Duo E8400 CPU @ 3 GHz and 3 GB RAM. Software profiling is carried out using the instrumentation profiling method, of the Visual Studio 2010 Ultimate Profiling Tool. This method provides detailed runtime data of every function including external function calls. Elapsed inclusive time values presented here show the time spent in the individual function and sub functions including time spend in calls to the operation.

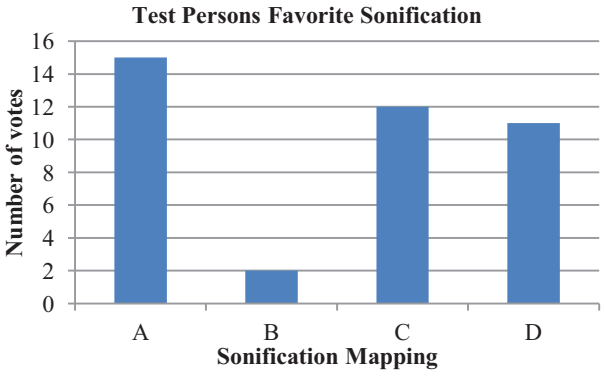


Figure 9: Test person's favorite sonification mapping

$$transmission\ time = \frac{message\ bytes * 9\ (\frac{bit}{byte})}{communication\ baudrate\ (\frac{bit}{s})} \quad (1)$$

In contrast to [5] the communication baud rate was increased to 460800 baud/s, in order to speed up the data transmission between Xsens bus master and computational hardware. The transmission time is calculated according to (1), according to the Xsens XM-B user manual. In sum the data generated per sampling instance consists of 81 bytes, comprising of 36 bytes per MTx sensor and a 7 byte preamble and 2 bytes for sample count. Compared to using a baud rate of 115200 baud/s this is a reduction of about 51 % by increasing baud rate. Xsens sensor system and Blueserial [22] Bluetooth adapters support this increased baud rate. Data acquisition and orientation computation lasts 2.55 ms in worst case. Therefore, sensor data transmission induced latency takes 4.44 ms.

The software caused latency is divided in the functional blocks for data processing according to [5]. Values listed in the Table 7 indicate the time per task to compute an update of the sonification parameters, comprising of enqueueing of sensor data items and calculation of the wrist position and sonification parameters. The usage of the ARM SIMD unit NEON, achieves a considerable latency reduction on the Integra processor for the floating point operation intensive computation of STK instrument generator audio samples, compared to the PC. The NEON unit achieves a speedup by computing up to 16 floating point operations in parallel. The NEON usage is activated by compiler flags. Data independent floating point multiplications are then computed in parallel.

A minimum audio buffer size of 150 audio samples is required, when operating using STK classes and the ALSA audio library. This results in a reduced latency, compared to the PC based approach where the Windows DirectSound library requires an audio buffer of at least 441 samples. In both cases audio buffer sizes below the mentioned limits result in an audio signal interrupted by clicking noise. Using an operation system like either Linux or Windows there is no way to directly access the audio device without using an audio buffer.

Software sub-block	Latency PC [ms]	Latency Integra [ms]
Fetch Data	$0.67*10^{-3}$	$27{,}90*10^{-3}$
Enqueue Data	$0.76*10^{-3}$	$0.70*10^{-3}$
Dequeue Data	$0.77*10^{-3}$	$0.70*10^{-3}$
Position Computation	$1.46*10^{-3}$	$4.20*10^{-3}$
Display movement features	$51.50*10^{-3}$	$49.50*10^{-3}$
Compute Sonification Parameters (sine)	$111.60*10^{-3}$	$68.53*10^{-3}$
Compute Sonification Parameters (beat)	$141.57*10^{-3}$	$100.09*10^{-3}$
Compute Sonification Parameters (instrument)	1.23	$150.14*10^{-3}$

Table 7: Detailed computational latency

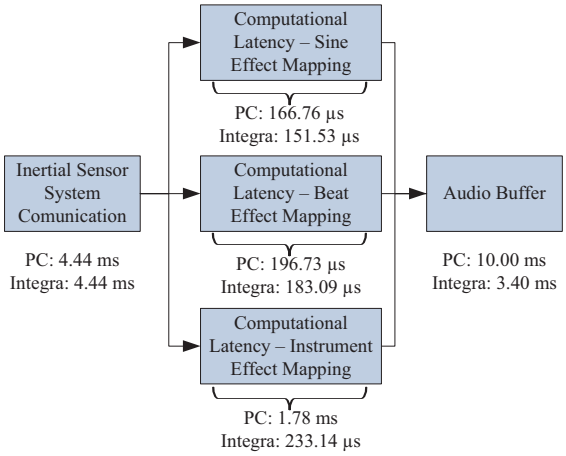


Figure 10: Hardware and software latency overview

The influence of data transmission, audio buffer size and computation, dependant on the hardware platform, is evaluated in Figure 10. In summary the overall system latency for the Integra processor sonification is about 8.07 ms, in contrast to a latency of 14.61 ms to 16.22 ms when operating on a PC. Major latency reduction is achieved by audio buffer minimization.

Figure 11 gives a comparison of computational costs of the required software tasks performed on PC and Integra platform. According to profiling the application allows a throughput of 4.28 kHz on the single core Cortex A8, as computation tasks last 233.14 µs at maximum. However, the maximum sampling frequency of the attached MTx sensor system will limit the application to an operating frequency of 100 Hz, when using two MTx sensors. Audio data rate was set to 44.1 kHz.

8. CONCLUSION

Implementing a mobile sonification system, the design goal is to achieve a sonification with an overall latency of 30 ms at maximum. The evaluation performed here clarifies that continuous, real time, low latency sonification of human arm movements can be achieved on low power, mobile platforms like the ARM Cortex A8 processor.

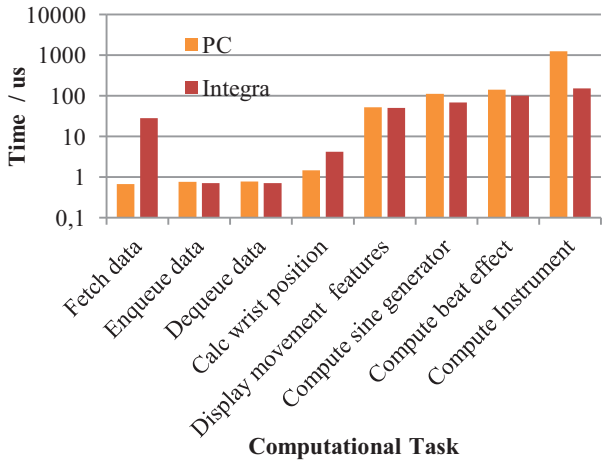


Figure 11: Computational latency distribution in comparison

Due to software optimization the overall computational latency keeps almost constant while performing with a significantly reduced clock frequency of 1 GHz compared to the 3 GHz PC.

Additionally, it is shown that depending on the operating system the audio buffer size can be significantly decreased. As the audio buffer size mainly influences the overall system latency this optimization step would also allow computing on processors with even lower clock rates and thus lower power consumption. Still the audio buffer causes one of the main latency parts. The second main inherent part is the MTx sensor data acquisition and data transmission time. In sum ≈98 % latency are caused by these two aspects.

In general, the overall latency of 7.99 ms of the proposed continuous sonification demonstrator meets the requirements and contains margin for operating it on platforms with further reduced clock rates and thus less power consumption.

The profiling results presented here also clarify, that more complex audio signal generation including mixing different fundamental or instrumental sound generation blocks would not significantly increase total latency. This enables further research in designing more comfortable and medical effective parameter mappings for audio synthesis.

The subjective test series performed here showed that all four evaluated parameter to sound mappings were significantly understandable. This is a convincing result, as none of the test persons had experience in designing or using movement sonification. All of the proposed mappings turned out to be intuitively usable, as the test persons had to rate the mappings after only 45 seconds of experience.

In overall rating, after performing free trials with all four sonification mappings, test persons rated the instrument based sonification to be best understandable. These mappings base on stereo effect in contrast to the beat effect in the competing two mappings. This shows that for unimpaired persons it is easy to correlate wrist position and sound source displacement.

In summary, the proposed Integra processor based system enables real-time low latency sonification. Additionally, it provides the required flexibility for adoptions in movement feature calculations and sound synthesis and enables further research in sonification design for upper arm movements. The hardware demonstrator will be used in studies to determine benefit from a continuous synthetic sonification in reach and grasp motor learning tasks. Studies will be used to figure out further significant motion parameters for relearning of movements and the design of an effective parameter to sound mappings, as well as an ambient and motivating sound design. The demonstrator is a research platform for designing a more effective and pleasant sonification for usage in home based stroke rehabilitation.

9. ADDITIONAL FILES

The attached “beat_sonification.wav” file represents an arm moving from the right to the front, then grasping a cup, moving it to the left and back to front. After that, the cup is raised for drinking and put back on to the table on the right. The file is available for download at http://www.ims.uni-hannover.de/fileadmin/www/files/forschung/sonification/beat_effect.wav

10. ACKNOWLEDGMENT

The work for this research project (W2-80118660) has been financially supported by the “Europäischer Fonds für regionale Entwicklung” (EFRE).

11. REFERNCES

[1] N. Schaffert, K. Mattes and A. Effenberg, "The sound of rowing stroke cycles as acoustic feedback," *The 17th Annual Conference on Auditory Display,* 2011.

[2] G. Kramer, "An introduction to auditory display", Addison Wesley Longman, 1992.

[3] Y. Tao and H. Hu, "3D arm motion tracking for home-based rehabilitation," *Proceedings of the 3rd Cambridge Workshop on Universal Access and Assistive Technology,* pp. 10-12, 2006.

[4] I. Wallis, T. Ingalls, T. Rikakis, L. Olson, Y. Chen, W. Xu and H. Sundaram, "Real-Time Sonification of Movement for an Immersive Stroke Rehabilitation Environment," *Proceedings of the 13th International Conference on Auditory Display,* 2007.

[5] H.-P. Brückner, C. Bartels and H. Blume, "PC-based real-time sonification of human motion captured by inertial sensors," *The 17th Annual Conference on Auditory Display,* 2011.

[6] T. Hermann, O. Höner and H. Ritter, "AcouMotion--An Interactive Sonification System for Acoustic Motion Control," *Gesture in Human-Computer Interaction and Simulation,* pp. 312-323, 2006.

[7] K. Vogt, D. Pirrò, I. Kobenz, R. Höldrich and G. Eckel, "PhysioSonic - Evaluated Movement Sonification as Auditory Feedback in Physiotherapy," *Auditory Display,* pp. 103-120, 2010.

[8] Texas Instruments, "C6-Integra™ DSP+ARM® Processor," [Online]. Available: www.ti.com. [Accessed 11 01 2012].

[9] Xsens Technologies BV, [Online]. Available: www.xsens.com. [Accessed 11 01 2012].

[10] G. Scavone and P. Cook, "RtMidi, RtAudio, and a synthesis toolkit (STK) update," *In Proceedings of the International Computer Music Conference,* 2005.

[11] J. McCartney, "SuperCollider, a new real time synthesis language," *Proceedings of the International Computer Conference,* pp. 257-258, 1996.

[12] I. Kosunen, K. Kuikkaniemi, T. Laitinen and M. Turpeinen, "Demonstration: Listen to Yourself and Others--Multiuser Mobile Biosignal Sonification Platform EMOListen," *Workshop on Multiuser and Social Biosignal Adaptive Games and Playful Applications,* 2010.

[13] V. Lantz and R. Murray-Smith, "Rhythmic interaction with a mobile device," *Proceedings of the third Nordic conference on Human-computer interaction,* pp. 97-100, 2004.

[14] M. Eriksson and R. Bresin, "Improving Running Mechanics by Use of Interactive Sonification," *Proceedings of the 3rd International Workshop on Interactive Sonification (ISon 2010),* 2010.

[15] G. Dubus and R. Bresin, "Sonification of sculler movements, development of preliminary methods," *Human Interaction with Auditory Displays--Proceedings of the Interactive Sonification Workshop,* pp. 39-43, 2010.

[16] Symbian OS, "Symbian smartphone operation system," [Online]. Available: http://www.symbianos.org/. [Accessed 01 11 2012].

[17] M. Fabiani, R. Bresin and G. Dubus, "Interactive sonification of expressive hand gestures on a handheld device," *Journal on Multimodal User Interfaces,* pp. 1-9, 2011.

[18] K. Vogt, D. Pirrò, I. Kobenz, R. Höldrich and G. Eckel, "PhysioSonic-Evaluated Movement Sonification as Auditory Feedback in Physiotherapy," *Auditory Display,* pp. 103-120, 2010.

[19] P. Maes, M. Leman and M. Lesaffre, "A model-based sonification system for directional movement behavior," *Interactive Sonification Workshop (ISon),* 2010.

[20] D. Levitin, K. MacLean, M. Mathews, L. Chu and E. Jensen, "The perception of cross-modal simultaneity," *International Journal of Computing Anticipatory Systems,* 2000.

[21] ARM, "ARM Cortex-A8 Processor," [Online]. Available: www.arm.com. [Accessed 11 01 2012].

[22] Hantz + Partner, "RS-232 Bluetooth Adapter," [Online]. Available: www.blueserial.de. [Accessed 11 01 2012].

[23] Nokia Corporation, [Online]. Available: http://qt.nokia.com/. [Accessed 11 01 2012].

[24] Texas Instruments, "TLV320AIC33/3106/34 Stereo Audio Converters," [Online]. Available: www.ti.com. [Accessed 11 01 2012].

[25] M. Hariri, M. Lakshmi, S. Larner and M. Connolly, "Auditory problems in elderly patients with stroke," *Age and ageing, Br Geriatrics Soc,* 1994.

[26] H.-P. Hesse, "Gehör: Psychoakustische und psychophysikalische Grundlagen", Kassel: Bärenreiter: Die Musik in Geschichte und Gegenwart, 2.Ausg., Bd. 3, pp.1104-1118., 1995 (In German).

[27] BS.1284-1, ITU-R Rec., "General methods for the subjective assessment of sound quality," 2003.

ACOUSTIC FEEDBACK TRAINING IN ADAPTIVE ROWING

Nina Schaffert

University of Hamburg,
Dept. of Human Movement Science,
Mollerstr. 2, 20148 Hamburg, Germany
nina.schaffert@uni-hamburg.de

Klaus Mattes

University of Hamburg,
Dept. of Human Movement Science,
Mollerstr. 2, 20148 Hamburg, Germany
klaus.mattes@uni-hamburg.de

ABSTRACT

Acoustic information contributes to the timing of human movements as sound conveys time-critical structures subliminally. That is of crucial importance for the technique training in high performance sports, where a successful movement execution depends on the precision of modifying the movement. Particularly adaptive athletes with visual impairments or blindness have a special sensitivity to acoustic information. Yet still only few sports can be practised by athletes with visual impairments.
Since a concept of providing online acoustic feedback during on-water rowing training sessions was introduced and empirically investigated with elite athletes, it was assumed that adaptive athletes particularly could benefit in terms of an enhanced perception for the movement execution. This paper deals with the implementation of providing online acoustic feedback to adaptive athletes in elite rowing. The results of the data-capture as well as the athletes' subjective experiences with the sound during rowing were described.

1. INTRODUCTION

The importance and relevance of sounds that accompany the execution of movements in sport situations is incontestable and is, among experts, a crucial criterion for the evaluation of the quality of a movement. Use of the sense of hearing to get a feeling for the movement is not a new approach in principle and it would be almost trivial to say that everyday movements (as well as sports actions) are always accompanied by sounds. The loudness of a sound event is the physical consequence of the kinetic energy of a movement. For experts, sound is equally as significant as the sensation, at the very least sounds play an important role in the feel for the movement, mostly without being explicitly obvious.
That said, auditory evaluation is an indicator or performance benchmark for the feeling of the movement, especially in situations in which the sense of hearing/auditory processing is prevented during the execution of movements. Only when the sound is missing does its essential importance for the movement execution become evident. In its absence, any feeling for the resulting forces and their effect on the movement is lost. In rowing, it is the sound of the boat's forward motion that provides the athletes with information about the boat velocity.
Sounds have a quite different and particular relevance for people with visual impairments or who are blind, and who have a special sensitivity to sound and tactile information due to their limited visual perception.
Despite the advances in technology up to the present, only few sports can be practised by athletes with visual impairments. With the help of additional provided acoustic information, it is possible for them to compensate for their deficiency in visual information-processing without being overloaded in terms of perceptual aspects. For example, in precision sports such as in elite biathlon, the most important success factor is in the capacity of alternating the skills of physical endurance and shooting accuracy during the competition. Athletes are assisted by acoustic signals, which depending on signal intensity indicate when the athlete is on target. Taking advantage of auditory perception, athletes fixate the target by ear. A bleep-beep tone represents the closeness to the centre of the target: the closer the aim at the centre, the shriller the tone. Another example for a non-visual sport game is the paralympic ball game Goalball that was created especially for blind people and athletes with visual impairment. In doing so, basic ideas were used such as a sounding ball [1]. These sports demonstrate impressively the adapted perceptual skills of sportsmen using non-visual information and the possibility of participating in sport (and even ball games) without any visual information. AcouMotion, a system for acoustic motion control, was developed by utilising existing technological possibilities to represent data acoustically and by integrating the method of interactive sonification, with a first application called Blindminton, a sports game similar to Badminton but designed for people with visual impairment [2]. The system presents information on the position of a virtual ball by using sound. Based on this information the player is expected to play a ball with a virtual racket against a wall without dropping it on the ground. This enables the presentation of auditory information in a more systematic way as in existing sport games using natural sounds such as the ringing of a bell inside the Goalball [3]. Furthermore, AcouMotion offers the opportunity to test audiomotor performance and specific performance-determining skills such as the auditory-perception orientation in space.
These systems open new pathways in high performance sports for visual impaired athletes. By use of the sense of hearing, it is possible to assess the surrounding situation [4]. Whereas attention can be focused on specific aspects of a sound source among a mixture of multiple, coexisting sound sources in order to extract the relevant information. A speciality of auditory perception is not to hear everything but 'to know' what needs to be heard and what needs to be paid special attention. This so called Cocktail-party effect [5] enables the listener to change the focus of attention from one sound source to another without

effort. In doing so, unimportant or disturbing noises or words (referring to conversation) are suppressed by focusing on the relevant information. Thus it is possible to perceive the interesting information as twice or three times as loud without turning the head [6].

Advantages of providing acoustic information about kinematic parameters in general as well as of the boat-acceleration time trace in particular have previously been described and empirically investigated in high performance rowing [7]. On the basis of these results and in order to support the feeling for the movement, as well as to provide an imagination of the duration of the movement and its execution for visual adaptive athletes, acoustic feedback is provided to elite adaptive athletes in on-water rowing training.

Special attention was paid to the effects subjectively perceived and athletes' reactions to the sound together with the results from data-capture, since even practising the sport is challenging for them. This is even more significant in situations with additional external influences during training such as the use of a test boat, measuring equipment and/or feedback-training methods. The use of synthetically produced acoustic information as a new training method is possibly even disturbing rather than beneficial for the execution of the rowing movement as athletes with visual impairment depend on auditory perception for their orientation. In comparison to sighted athletes it is not possible for them to subordinate the auditory sense.

This paper describes the results of providing acoustic feedback online during on-water training to adaptive athletes in elite rowing and their experiences subjectively perceived via the sound during rowing.

2. METHODS

2.1. Adaptive rowing

The regulations for adaptive rowing require that the crewmembers must have a handicap. In boat classes for more than two athletes, the crew must, more specifically, consist of athletes who are physically disabled as well as visually handicapped (part or blind). The crew studied consisted of two visual impaired athletes, one of whom was blind (100%) as well as of two physically handicapped athletes. The exceptional challenge for the blind athlete was his lack of rowing experience in terms of a perception as well as of a feeling for the rowing movement. The primary aim during the preparation phase for the adaptive world championships was set on synchronising the crew in a uniform rhythm in order to qualify for the Paralympic Games 2012 in London.

2.2. Characterization of the rowing stroke cycle

The rowing stroke is a cyclic motion sequence, separated into two main phases drive and recovery (or release), which are further subdivided into the front and back reversal (also known as the catch and finish turning points). With regards to the boat acceleration-time trace, the rowing cycle begins with minimal acceleration followed by a distinctive increase during the catch and the drive phase to the point of maximum boat acceleration. The end of the drive phase is represented by the next local minimum in acceleration. It is the transition phase where the oars were lifted out of the water (back reversal). The recovery phase begins subsequently to the transition phase with minimal acceleration amounts and ends a global minimum in acceleration. It is subdivided into a first and a second phase. The classification of the several phases in the rowing cycle is made in relation to a description of the rowing movement as well as to the executed technical skills.

The primary and overwhelming importance of the recovery phase with regards to the propulsive effect of the rowing cycle becomes manifestly clear. At the end of the drive phase, when the blades emerge from the water, the boat is released to run forward. This movement is challenging for the athletes after raising the oars out of the water, as they have to glide back up to the catch again in order to prepare the next stroke. Thus, it is important to execute the recovery phase without reversing the boat's momentum, that is, athletes' mass must be carefully moved by sliding towards the stern. This phase is critical for the boat velocity in particular, because fluctuations occur as a result of energy dissipations by jerky movements. Consequently, athletes should integrate the several parts of the rowing stroke into one movement that is as consistent and smooth as possible. This is especially important because one movement phase flows into the next one. However, when rowing at higher stroke rates it is not possible to strictly separate the single movement phases from each other.

2.3. Subjects

The athletes participating in the study were members of the German national adaptive rowing team (N=6), male (n=3) and female (n=3). The coxed four (LTA4+) was accompanied during on-water training sessions for two weeks and over a total of seven training sessions. For several reasons, it was not possible to train with the original crew for the whole time and so several times substitutes sat in four and came into contact with the sonification.

2.4. Measurement System

The acoustic feedback system *Sofirow* [8] (developed in cooperation with engineers from BeSB GmbH, sound and vibration, Berlin) [9] was used. The device measured the kinematic parameters: propulsive boat acceleration (a_B) with a micro-electro-mechanical (MEMS) acceleration sensor (sampling rate adjustable up to 125Hz) and boat velocity (v_B) with GPS (4Hz). Figure 1 showed the system and its position location on top of the boat.

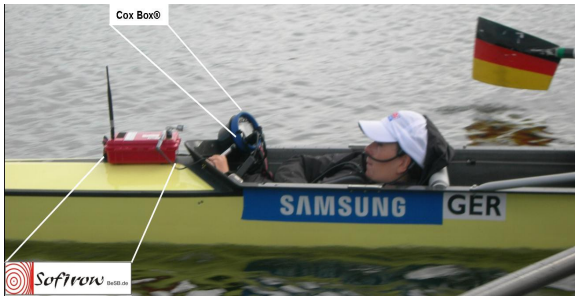


Figure 1: The acoustic feedback system *Sofirow*.

Sofirow converted the boat's acceleration-time trace online into acoustic information and transmitted the sound sequence via WLAN to the athletes in the rowing boat as well as to the coach into the motor boat. The sonification was presented in addition to the natural soundscape via loudspeakers of the inboard existing Cox Box® (Nielsen-Kellermann) through which the coxswain as well as the coach communicates instructions to the crew. Thus it was possible to listen to the sonification, to the coxswain as well as to the coach at the same time. In order to control the timing and the duration of the acoustic feedback, the sound could be selectively switched on or off by remote-control from the accompanying coaching boat.

In doing so, it was possible for the coach and the scientist to listen to the sonification while the athletes did not receive the acoustic feedback. Acoustic transmission was controlled by the scientist agreeing with the coach listening to the same acoustic feedback simultaneously with the athletes or alone.

The data storage on a SD-card made it possible to analyze the effect of the acoustic feedback on the boat motion in real time as well as to re-sonify the data subsequently.

2.5. Sound Design

The data-to-sound-transformation was achieved with the software Pure Data (Pd) as previously described and established in an earlier investigation with the German national rowing team in on-water training sessions. Using the sonification technique of Parameter Mapping [10], the boat's acceleration-time-trace was directly mapped to tones on the MIDI-scale and related to tone-pitch. In doing so, the data were transformed algorithmically into an audible sound in real time as a direct modulation. Consequently, tone pitch changed as a function of the boat's acceleration-time-trace and represented and differentiated between qualitative changes in the boat motion.

2.6. Test Design and statistical analysis

The investigation took place at the race course in Ratzeburg, Germany in August 2011 during the preparation phase for the adaptive world championships in Bled, Slovenia.

Prior to the first on-water training session, the athletes were introduced to the sonification in order to give them an idea of what they have to expect. Therefore, the sound sequence of a stored training run which was synchronized with a video was presented to the athletes. They could listen/watch to it as often as they needed.

The presentation of the acoustic feedback during on-water training was adjusted accordingly to the special needs of the athletes with visual impairments without overloading their environmental perception. Thus, the acoustic feedback was presented in up to 3 blocks per training session and for a total of 12 blocks. Each block consisted of 4 sections without and with the presentation of acoustic feedback in alternating order for the duration of 500m respectively.

In order to conduct an online analysis, the scientist and the coach listened to the sound result in the motorboat while the athletes did not receive any feedback. For the analysis, the sections were separated, consisting of a total of 30 rowing cycles each rowed at a comparable stroke frequency (±0.5 strokes per minute) for all sections.

Statistical comparison was achieved using an ANOVA (general linear model) with repeated measures (level of statistical significance was set at $p<0.05$) with the software SPSS 16.0. This procedure allows the test of interdependencies as well as of impacts (effects) from single factors between the sections studied. In order to rate the size of one factor or combination of factors, partial eta-squared (η^2_p) was calculated as the parameter of effect size. Partial eta-squared describes the effect size on the dependent variables according to the classification according to Cohen [11]. Post-hoc tests were used to rate the differences between the sections studied by comparing them pairwise. The statistical analysis considered the sections without and with acoustic feedback (AF), labeled as follows:

Baseline reference section (without acoustic feedback)
Section 1 (with AF)
Section 2 (without AF)
Section 3 (with AF)

Standardized questionnaires were taken in addition to examine the perception of adaptive athletes of the acoustic feedback in terms of its comprehensibility, correspondence with the rowing movement, its attention-guidance function for specific movement sections as well as potentially disturbing aspects.

3. RESULTS

The results of the investigation were described in separated subsections as follows: data-capture (3.1) describes the effects of acoustic feedback on the mean boat velocity; questionnaire (3.2) describes athletes' reactions to the sound and the effects subjectively perceived.

3.1. Data-capture

The results of the sections with acoustic feedback show that at training stroke frequency (SF 20 +/- 0.5 strokes per minute) there is a significantly increased mean boat velocity for the sections with sonification in comparison with the baseline (reference/control section) without sonification ($F_3=3.79$; $p=0.03$; $\eta_p^2=0.35$). The value for the effect size (partial eta-square) shows mid-level effect power.

According to the coach's GPS, the "*sections with the sonification were* (...) *faster with the sound*" and the crew "*moved away from the motorboat*". In particular, in the first section with sonification the mean boat velocity with acoustic feedback was increased ("*more clearly and better*"). In the subsequent sections without, with and without sonification the increases were less emphasized; the cox stated that the stroke frequency was however slightly increased ("*a frequency of 20 was more easily maintained with acoustic feedback than without. In the sections without the sonification it rose to 21 more often than with tone*.") With more training sessions in which the sonification was introduced, it was clear that the athletes were better able to achieve the increases in speed at a constant stroke frequency in the sections without sonification. The changes were most clear at the front reversal, which is represented in the sound sequence by a deep tone. With a

movement executed too slowly, the tone is momentarily inaudible. The aim was to reduce the duration of the reversal movement by means of an uninterrupted sound-sequence.

In order to rate the difference between the sections studied, the results for the pairwise comparisons were considered. During both sections with acoustic feedback (section 1 and 3), the mean boat velocity increased significantly compared to the baseline (reference section) without. In contrast, the section without acoustic feedback (section 2) showed no significant differences to the baseline. The values for partial eta-square show high-level effect power for both sections with acoustic feedback (section 1 and 3) and mid-level effect power for the section without (section 2) (table 1).

Table 1: Test of contrasts (within-subjects) for the effect of acoustic feedback on the boat velocity in the different sections studied vs. the baseline: F-value (F), level of significance (p) and partial eta-square (η_p^2); degree of freedom=1; *N*=12.

Sections		F_1	P	η_p^2
	s1 (with)	10.33	0.01	0.60
Baseline	s2 (without)	2.88	0.13	0.29
	s3 (with)	7.42	0.03	0.51

Figure 2 provides a visual impression of the differences measured between the sections with and without acoustic feedback in comparison to the baseline. As demonstrated in the figure (2), the sections with acoustic feedback showed a distinct increase in the mean boat velocity.

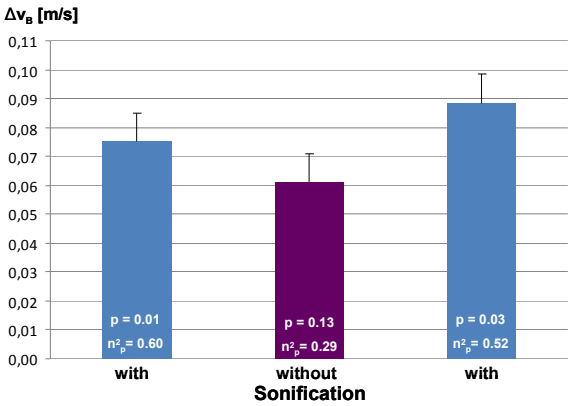


Figure 2: Mean differences and standard errors for the boat velocity (Δv_B) for the sections studied in comparison to the baseline.

3.2. Questionnaire

In reply to the question of what the athletes changed technically in terms of the movement execution, "*sliding forward gently*" and "*catches*" were emphasized and it was stated that they tried to keep the tone "*as constant as possible*" and "*maintained as long as possible before the entry of the blades*". For the coach, the efforts of the athletes were clear ("*the movement seemed smoother*" and that "*they could carry out the front reversal pretty well*").

The perception of the adaptive athletes when rowing with the sonification was "*initially irritating*" and "*confusing*", since the "*unusual type of training up to the present was not in use in normal training*". "*It had taken some time before I was used to the tone* (...) *until it had become a part of you*" (...) "*and after a familiarisation phase it was possible for us to improve the run of the boat*." And "*as long as the toning does not overwhelm the background sounds which are for me important, it is helpful, and enables an improved check on the individual rowing technique*," as well as a "*focussed improvement of the weak points in the movement*." "*If I don't want to hear it, I can 'blank out' the tone*." "*With regulated use in training, the toning is good*" but "*one must first learn to 'hear' it*." "*If I can concentrate fully on the toning*" variations between individual strokes become clear. "*Extremely good for the forward sliding and fast catches or variations between them: absent tone with too slow catches/releases*."

The procedure for the presentation of acoustic feedback was adjusted to the special need of adaptive athletes. This was confirmed with athletes' statements who appreciated the way of presentation as appropriately for on-water training sessions: "*In that way it was practised: very well for 500m-sections with and without tones in alternating order. It might be well to have two days of rest between the training sessions with the sonification*." In order to be helpful the sonification should be used regulated in the training session because "*(...) if it is used overmuch, (...) the tone becomes annoying*". Here, too, it became evident that athletes do have individual strategies in dealing with the sound. "*I could have listen to it more times in order to fix it on my mind*" (…) "*for me as a sighted athlete, my own feeling for the boat run and the movement execution is more important*" (…) "*I would like to test it in the single sculls*".

The results underline previous findings and give support to our initial assumptions that acoustic feedback provides assistance for adaptive athletes to enhance their perception for executing the rowing movement more effective.

4. DISCUSSION

This paper described the results of providing acoustic feedback online during on-water training to adaptive athletes in elite rowing and their experiences subjectively perceived via the sound during rowing. It was aimed at enhancing athletes' perception for movement execution with the final aim to synchronise the crew in a uniform rhythm in order to improve the boat velocity by a reduction of intracyclic interruptions in the boat acceleration.

A theoretical basis for this concept as well as a design for a rowing specific acoustic feedback system has previously been described and empirically investigated with the German national rowing team [7]. With *Sofirow*, an acoustic feedback system, it is possible to provide the rhythm of the rowing cycle audibly by sonifying the boat acceleration-time trace. In doing so, changes in the measured acceleration trace were correlated to tone pitch: with increasing boat acceleration, the sound sequence increased in terms of tone pitch. Changes, that are normally invisible by watching the boat traveling through the water, became evident, as the differences were tiny but affect the boat motion importantly.

The acoustic feedback reflected overall effects of all external forces (water resistance, etc.) as well as athletes' movements acting on the system as a whole (boat and rower) by providing the boat acceleration time trace audibly. Athletes perceived the sound information of the movement patterns independently from vision and thus, the medium of presentation was supportive and enhanced their perception of the boat run. Interactivity of the perception process was allowed within the time frame of neuronal information acquisition and processing [12], and, as a result, the control of executing the movement was realizable in a time-uncritical way. In contrast to the coach's verbal instructions that sometimes need further explanations, the sound result was intelligible to all. Thus, the psychological interaction between the coach and athletes was bridged.

Owing to the direct coupling of tone pitch to changes in the boat's acceleration-time trace, the information contained in the captured-data became intelligible for the athletes, directly and intuitively and athletes perceived the single rowing cycle as a short sound sequence. Periodic recurrence of characteristic sections inside the rowing cycle represented the rhythm of the rowing cycle and awakened sensitivity for details in the sequence without further explanations needed. Awareness of the structure emerged solely from the knowledge of the movement and audio-visual interaction [13]. Rhythm is defined in movement science as a temporarily sequence of motor actions whose timing is of crucial importance for the movement execution [14]. It thus is inseparable from synchronization within moving contexts. Consequently, it was assumed that the measured improvement in the mean boat velocity was due to both, improved crew synchronization as well as due to improvement of the individual rowing technique of the athletes. This was confirmed due to athletes' individual statements.

With that, the results are similar in principle to previous findings in elite rowing training conducted with sighted and physical not handicapped athletes. Using the sonification as a new feedback method in the technique training of adaptive elite athletes in the four (LTA4+), it was possible to give support to the creation of an imagination of the movement as well as to the feeling for the rowing movement. The excited and keen interest of the coach and the crew in the sonification and its implementation into the technique training is promising for a regularly use in on-water training of adaptive athletes with the potential to expand it to other crews with a handicap. The feedback-training method will be an integral part for the preparation for the London 2012 Paralympic Games.

5. CONCLUSIONS

The acoustic feedback system *Sofirow* was developed to support the control of the rowing movement by the presentation of information that is provided through the sense of hearing in addition to existing sensory channels such as the visual sense. Thus, the device complements the feedback training in addition to existing feedback systems and provides relevant information for athletes with visual impairments. Captured sonified data of the boat's acceleration-time trace are stored as audio files (wav file format) and available for mental training. It furthermore contributes to previous research in rowing biomechanics [15], [16], [17] [18] and complements the existing visual analysis of the rowing technique used for the biomechanical diagnostic [19] with an expansion for the audible domain.

Desirable for the future is the willingness of the German Rowing Association (DRV) for funding adaptive athletes which is still reserved.

6. ACKNOWLEDGMENTS

We would first like to thank the coach Thomas Boehme and the athletes of the German adaptive national rowing team as well as the German Rowing Association (DRV) and the National Paralympic Committee Germany (DBS) for their cooperation. Congratulations for the qualification for the London 2012 Paralympic Games.

Many thanks as well go to the engineers of BeSB GmbH Berlin Dipl.-Ing. Reiner Gehret and Sebastian Schroeter for developing the acoustic feedback system *Sofirow* and the technical support. We also want to thank Bruce Grainger for critically proof reading the manuscript.

7. REFERENCES

[1] O. Hoener & T. Hermann, *Listen to the ball! – Sonification-based sports games for people with visual impairment*, Bielefeld University, Germany 2005

[2] T. Hermann, O. Hoener. & H. Ritter, "AcouMotion – An Interactive Sonification System for Acoustic Motion Control." In S. Gibet, N. Courty & J.-F. Kamp (Eds.), *Lecture Notes in Artificial Intelligence 3881* (pp. 312–323). Berlin, Heidelberg: 2006.

[3] Goalball: http://www.ibsa.es/eng/deportes/goalball/presentacion.htm; last time retrieved: Feb. 27th, 2012.

[4] A. S. Bregman, "Auditory scene analysis: Hearing in complex environments", in S. McAdams and E. Brigand (Eds.), *Thinking in sound: The cognitive psychology of human audition.* New York: Clarendon Press/Oxford University Press, 10-36, 1990.

[5] E.C. Cherry, "Some experiments on the recognition of speech, with one and with two ears." *Journal of the Acoustical Society of America* 25, 975-979, 1953.

[6] N. Birbaumer & R.F. Schmidt, *Biologische Psychologie.* Springer Medizin Verlag. Heidelberg, 7. Auflage, 2010.

[7] N. Schaffert, K. Mattes. & A.O. Effenberg,. „An investigation of online acoustic information for elite rowers in on-water training conditions." *J. Hum. Sport Exerc*. 6 (2):392-405, 2011.

[8] *Sofirow*: www.sofirow.com; last time retrieved May 8th, 2012.

[9] N. Schaffert. & K. Mattes, "Designing an acoustic feedback system for on-water rowing training." *Int J. Comp* Sci Sport, 10 (2): 71-76, 2011.

[10] T. Hermann, "Taxonomy and Definitions for Sonification and Auditory Display", in *Proc. 14th Int. Conference on Auditory Display (ICAD)*, June 24-27, Paris, France, 2008.

[11] J. Cohen, *Statistical power analysis for the behavioral sciences*. Hillsdale: Lawrence Erlbaum, 1988.

[12] H. De Marées, *Sportphysiologie*, Köln Verlag Sport und Buch Strauß, 2003.

[13] A. O. Effenberg, "Multimodal Convergent Information

Enhances Perception Accuracy of Human Movement Patterns", in *Proc. 6th Ann. Congress of the European College of Sport Science (ECSS)*, Sport und Buch, Strauss, 122, 2001.

[14] W. Auhagen, „Rhythmus und Timing", in H. Bruhn, R. Kopiez and A. C. Lehmann (Eds.), *Musikpsychologie. Das neue Handbuch*, Rowohlt Verlag, Reinbek bei Hamburg, 437-457, 2008.

[15] K. Affeld, K. Schichl & A. Ziemann, „Assessment of rowing efficiency". Int J Sports Med; 14 Suppl 1, 39-41, 1993.

[16] M. McBride, "Rowing Biomechanics." In V. Nolte, (Eds.) *Rowing faster*. Human Kinetics Publishers, Inc., 111-124, 2005.

[17] V. Nolte, (Ed.) *Rowing faster. Serious Training for serious rowers*. 2nd edition. Human Kinetics, 2011.

[18] Böhmert, W. and Mattes, K. 2003. Biomechanische Objektivierung der Ruderbewegung im Rennboot. In Fritsch, W. (Eds.). Rudern - erfahren, erkunden, erforschen. Gießen: Wirth-Verlag (Sport Media), 163-172.

[19] K. Mattes & W. Böhmert, "Biomechanisch gestütztes Feedbacktraining im Rennboot mit dem „Processor Coach System-3" (PCS-3)." In J. Krug & H.-J. Minow (Eds.). *Sportliche Leistung und Techniktraining*. 1. Gemeinsames Symposium der dvs-Sektionen Biomechanik, Sportmotorik und Trainingswissenschaft vom 28.-30.9.1994 in Leipzig, St Augustin: Academia, 283-286, 1995.

PERCEPTUAL EFFECTS OF AUDITORY INFORMATION ABOUT OWN AND OTHER MOVEMENTS

Gerd Schmitz

Leibniz University Hannover,
Institute of Sport Science,
Am Moritzwinkel 6, 30459 Hannover, Germany
gerd.schmitz@sportwiss.uni-hannover.de

Alfred O. Effenberg

Leibniz University Hannover,
Institute of Sport Science,
Am Moritzwinkel 6, 30459 Hannover, Germany
alfred.effenberg@sportwiss.uni-hannover.de

ABSTRACT

In sport accurate predictions of other persons' movements are essential. Former studies have shown that predictions can be enhanced by mapping movements onto sound (sonification) and providing audiovisual feedback [1]. The present study investigated behavioral mechanisms of movement sonification and scrutinized whether effects of own movements and those of other persons can be predicted just by listening to them. Eight athletes heard sonifications of an indoor rower and quantified resulting velocities of a virtual boat. Although boat velocity was not mapped onto sound directly, it explained subjects' quantifications by regression analysis (R-squared = 0.80) significantly better than the directly sonified amplitude and force parameters. Thus perception of boat velocity might have emerged from those sonifications. Predictions of effects of unknown movements were above chance level and as good as predictions of own movements. Furthermore athletes were able to identify their own technique among others (d' = 0.47 ± 0.43). The results confirm large perceptual effects of auditory feedback and - most importantly - suggest that movement sonification can address central motor representations just by listening to it. Therefore not only predictability but also synchronization with other persons' movements might be supported.

1. INTRODUCTION

Transforming human motion into sound has been the exclusive domain of musicians. But sonification of human movement data has proved to support perception and action in sport: sonifying the ground reaction force of counter movement jumps enhances the perceptual accuracy of jump height ratings, and results in enhanced movement performance, when jumps are reproduced [1]. Although there is growing evidence for the efficacy of sonification, the underlying mechanisms are largely unknown. One possible mechanism is a co-activation of auditory and motor areas in the brain: the listening to a piano melody activates motor areas in the brain, when this melody has been practiced for just 30 minutes [2]. Another mechanism might be enhanced activation of multimodal brain areas: Using the same stimuli as Effenberg [1], Scheef et al. [3] found increased neuronal activation in multimodal brain areas for audiovisual congruent compared to incongruent stimuli, suggesting an amplifier effect of sonification on motor perception. But further mechanisms are probable. A key player for the understanding of other persons' actions is the human action observation system: This systems harbors the so-called mirror neurons that are activated when a person performs an action or when this person observes another person performing the same action [4]. Knowledge of the mirror neuron system comes from studies with visual stimuli, but two recent studies suggest that natural sounds and music address the mirror neuron system as well [5,6]. Since this system is active during the observation of other persons' actions as well as when movements are preformed, it might be the neural interface between perception and action. The hypothesis is that during action observation the mirror neuron system activates the own motor system to internally simulate the movement and its outcome. In consequence predictions should be more accurate, the higher the individual motor experience in the observed task is, and experts should predict outcomes of sport-specific movements better than novices. Actually a study from Aglioti et al. [7] suggested that this is an effect of motor experience on perceptual accuracy: when basketball players, trainers and journalists have to predict the outcome of free shots at the basket, players perform best.

If motor experience shapes perceptual accuracy, effects should not be limited to sport-experts only, because everybody is expert of his own individual movements. Therefore everybody should predict actions best, when he or she observes his own actions ("own-effect"). Several studies have investigated this hypothesis using visual stimuli and found small but significant effects: when dart throws or handwriting strokes had to be predicted, predictions were most successful when the effect of the own movements - and not of movements from other persons - were observed [8,9].

Prediction and identification of actions might not depend on holistic and natural presentations of bodies. Former studies have shown that it is sufficient to display the large joints as point-lights [8,10]. But it still remains unclear which movement parameters provide relevant information. The results of Loula et al. [11], who reported different identification rates for dancing and boxing compared to walking and running, suggest that the significance of parameters varies between movement categories. Therefore a detailed investigation of this aspect is reasonable.

The cited studies argue for a close relation between action and visual perception, notably an internal simulation of movements by the own motor system, when actions are observed. One study reports a similar effect from the field of

This work is part of the project "Kognition in Bewegung" (WIF 60460288) at the Leibniz University Hannover.

music: Keller et al. [12] found that pianists synchronize their movements better with recordings of their own than with recordings of other persons, indicating that the "own-effect" might not be limited to the visual domain. The demonstration of an "own-effect" for sonification would broaden the knowledge about behavioral aspects and their neural mechanisms addressed by sonification and about motor representations: providing evidence for an internal action simulation on the basis of sonification would suggest that motor representations are multisensory. Therefore one goal of the present study was to investigate, whether sonified movements are anticipated best when they are own movements, and if own movements can be identified by their sonification.

In addition to the theoretical knowledge sport practical implications can be expected: Movement coordination and synchronization depend on action prediction. This is a common principle for intraindividual synchronizations as the coupling of hands [13], as well as the interindividual synchronization of two or more persons [14]. Therefore any team sport and many forms of social interactions should benefit from optimized predictability of own and other movements.

Predictability does not depend on motor experience alone: a crucial factor is the accurate perception of significant movement parameters. Since there is an overflow of information into our sensory systems we have to focus our attention onto single parameters and in this way filter information streams. Years of sport-specific training are necessary to develop perceptual expertise and to direct the attention to important and neglect unimportant movement parameters. Therefore sport-experts show improved perceptual performance compared to novices and predict movements better [15]. In addition to the expertise effect predictability of movements can be enhanced by other mechanisms: Team players often exaggerate their own movements to make them perceivable and predictable to their team mates [16]. Movement sonification can address these issues twofold: 1. Attention can be focused more easily when relevant parameters are accentuated by sonification. But this requires the knowledge of the relevance of parameters. 2. The continuous mapping of movement parameters onto sound enhances the perceptual accuracy in observers, since it provides complementary information to the visual and kinesthetic modality, yielding superadditive integration effects [3], as well as additional or accentuated information about movement features. Therefore a second goal of the present study was to analyze which parameters among others are chosen by athletes to predict action effects and to identify the own movement.

2. METHODS

Eight rowing athletes (21.8 ± 9.2 years) participated in the study. They all had been nominated by the state coach due of their high technical qualification. In a first session they performed 50 minutes on an indoor rower (Concept2, Inc., VT, USA). After 5 minutes of rowing at a self-chosen velocity they were instructed to follow eight different velocities in three blocks of 15 minutes, interleaved by rest breaks of about 10 minutes. Two types of real-time feedback were provided to the athletes: A) Virtual boat velocity was calculated online and displayed by the indoor rower itself, permitting target-performance comparisons. All athletes were familiar with this kind of feedback from their own training. B) Most importantly athletes heard a sonification of their rowing performance via earphones (AKG K330). An exemplary stimulus is attached as supplementary file. The sonification system was described in detail previously [17] and only the main elements will be reported here: The indoor rower was featured with two incremental encoders and two force-sensors attached to the handle, seat and foot rest, measuring grip force and amplitude, seat amplitude and foot rest force (sampling rate 100 Hz, FES Berlin ®). Movement parameters were mapped onto sound using standardized MIDI control messages [18]. Parameter variations were linearly (kinematics) or non-linearly (dynamics) proportionally to modulations of pitch and loudness. Mapping characteristics were standardized inter-individually.

Sonification of four parameters is characterized by a high information density. In addition to the magnitude of the two kinematic and two dynamic parameters, it informs about temporal aspects of the movement: It could be possible to perceive movement frequency by identifying the frequency of similar sound patterns (for example detection of the absolute minimum of the grip amplitude, Figure 1). Combining those information then might built further percepts of mechanical power or individual technical patterns.

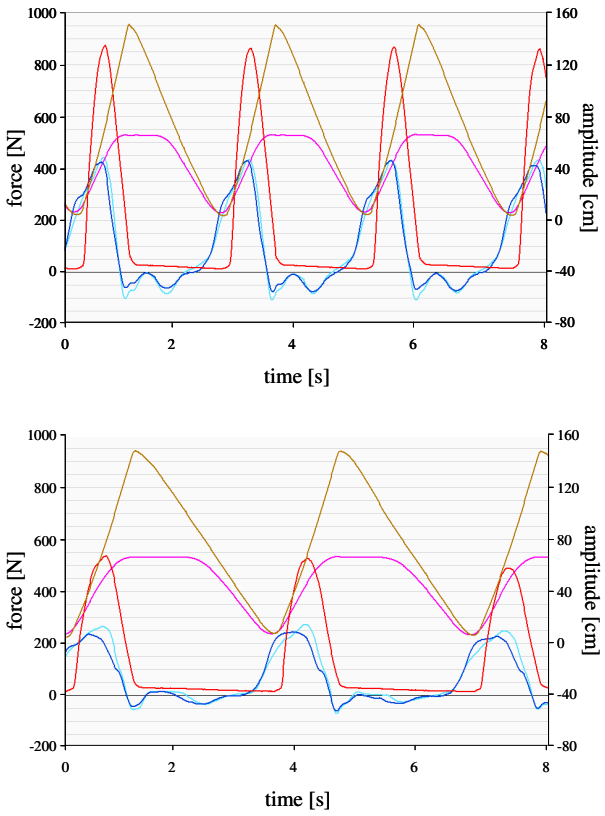


Figure 1: Grip amplitude (light red), seat amplitude (pink), grip force (dark red) and foot rest forces of the left (light blue) and right foot (dark blue) during slow (top) and fast (bottom) rowing cycles.

Perceptual effects of the sonification were investigated nine to twelve days after the rowing session. Each athlete heard sonifications of his *own*, of a person *known* from training or *unknown*. This design was created in accordance to Loula et al. [11], who reported higher identification rates for own movements than for movements of known and unknown persons, confirming the above mentioned "own-effect". Furthermore identifications were better when movements of *known* persons were observed than those of *unknown* persons, which can be interpreted as significant influence of perceptual expertize on movement perception.

One trial consisted of two consecutive stimuli. Length of stimuli varied randomly and contained about two rowing cycles. Stimuli of one trial were from the same person (*own, known, unknown_same*) or from two different persons (*unknown_different*). 30 trials of each treatment were presented to the athletes yielding 120 trials in one session, arranged pseudo-randomly. Before the session started, subjects received in three trials knowledge of results. This procedure was repeated every 30 trials.

Athletes were instructed to (1.) quantify differences of virtual boat velocities within one trial (task 1: 120 estimations), differing within a range of ±1.4 m/s and (2.) to detect own techniques from the sonifications (task 2: 240 decisions). Virtual boat velocity v [m/s] was calculated on the basis of the mechanical power P [W] at the grip as

$$v = 500m * \left(\frac{dF * 10^6}{P} \right)^{-\frac{4}{11}} \quad (1).$$

(dF - the drag factor of the wind wheel, which depends on the position of wind panel - was inter-individually standardized at 125 Nms^2). This velocity matches the virtual boat velocity calculated by the indoor rower itself.

3. RESULTS

All subjects performed well at all velocity stages in session I. Movements of different velocities of a single subject are illustrated in Figure 1. Parameters varied marginally between subsequent cycles, indicating that indoor rowing performance was highly stereotyped (Figure 1). Therefore sonification of those parameters resulted in highly stereotyped sounds that were provided to the subjects in real-time.

3.1. Velocity estimations

The perceptual effect of this sonification was investigated in task 1, when subjects quantified velocity differences of two rowers. Velocities of the virtual boats differed from -30% to +40% and subjects´ estimations filled the complete spectrum (Figure 2). To evaluate if subjects had followed the experimenter's instructions and based their estimations on evaluations of the virtual boat velocity, it was analyzed whether subjects' estimations could be best explained by the complex parameter virtual boat velocity – not directly perceivable - or other parameters as grip force maximum, foot rest force maximum, grip amplitude and seat amplitude, which could directly be perceived via pitch and loudness differences. Linear regression analysis yielded best predictability of subjects' estimations by virtual boat velocity, explaining 80% of variance ($F(1,955)=3926.55$, $p<0.001$). Significantly less variability ($t(954)=13.38$, $p<0.001$) was explained by the force maxima (grip force: R-squared=0.67, $F(1,955)=1982.66$, $p<0.001$; foot rest force: R-squared=0.66, $F(1,955)=1821.15$, $p<0.001$) and marginal or no correlations were evident for grip amplitude (R-squared=0.01, $F(1,955)=13.93$, $p<0.001$), and seat amplitude (R-squared<0.01, $F(1,955)=1.72$, $p>0.05$). Therefore perceptual results are best described by virtual boat velocity. Most importantly, explanation of 80% of variance means that only 20% of variability are due to individual differences and preferences, biases and random errors (Figure 2).

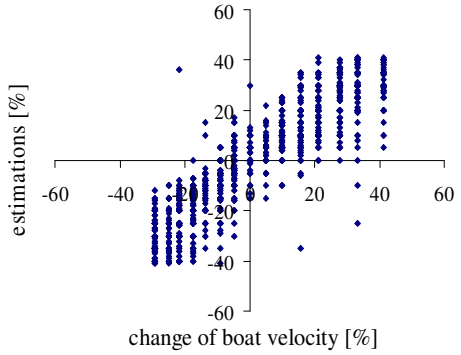


Figure 2: Correlation of estimated and calculated virtual boat velocity of an indoor rower.

To analyze perceptual accuracy to own or others´ sonifications the absolute error between estimated and given change of boat velocity was calculated for the different treatments. Figure 3 illustrates across-subjects´ means and standard deviations: Absolute errors were significantly below chance level ($t(7)=-24.09$, $p<0.001$), which was defined as absolute error of constant estimations of 0% velocity difference. Results differed between treatments as confirmed by one-way analysis of variance ($F(3,21)=4.10$, $p<0.05$).

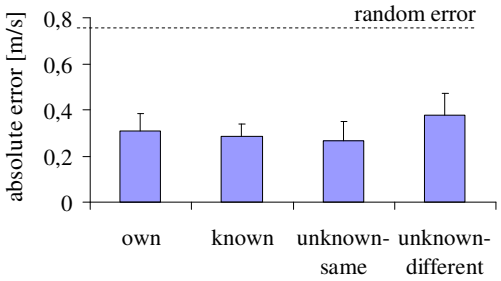


Figure 3: Absolute error [m/s] between estimated and absolute difference of virtual boat velocity when listening to sonifications of the *own* technique, technique of *known* or *unknown* persons.

Decomposing this effect by Scheffe's post hoc test yielded no differences between *own*, *known* and *unknown_same* (all $p>0.05$). But estimations were better

($p<0.05$) when subjects subsequently heard sonifications of the same rower (*unknown_same*) than of two different rowers (*unknown_different*). The results indicate a high perceptual performance, but performance was not better when a subject heard his own sonification.

3.2. Identification

Task 2 was to explicitly judge whether the provided sonifications were from the own or from other persons´ techniques. Subjects correctly identified their own rowing in 40 ± 16% of all cases, which is significantly above chance level of 25% ($t(7)=2.500$, $p<0.05$). They correctly rejected their own technique to 76 ± 12%, which is close to chance level of 75% ($t(7)=0.289$, $p>0.05$). It should be considered that identification rate could be positively biased by the tendency to identify a technique as "own" or negatively biased by the tendency to identify a technique as "not own". Subjects of the present study responded in 28 ± 11% of all trials that they had heard their own technique, a value that nearly matches the correct rate of 25% ($t(7)=0.715$, $p>0.05$). Nevertheless, response biases might have influenced the results and should be eliminated from analysis. A common procedure is to calculate the discrimination index *d'* as unbiased identification variable, that considers individual relations of *hit* rates (correctly identifying the own technique) and *false alarm* rates (wrongly identifying a technique as "own") [19]. Subjects of the present study yielded a *d'* of 0.47 ± 0.43, which is significantly larger than zero ($t(7)=3.10$, $p<0.05$), confirming a significant detection of own among other techniques.

To scrutinize if identifications can be ascribed to one or more movement parameters exploratory discriminant analysis were calculated. In addition to the four sonified parameters, two technique-related parameters were included as predictors. An initial impulse can be optimized when the grip force reaches its maximum early in time. Therefore t_*grip* was calculated as time of maximal grip force in relation to the duration of the rowing cycle. Impulse transmission from foot rest to grip force necessitates temporal coupling of both forces, which can be expressed by the quotient of the points in time of both force maxima (t_*grip/footrest*). The optimal coupling of both force maxima depends on the anthropometry of the athlete and therefore differs inter-individually; thus each athlete might have his own optimal value and t_*grip/footrest* might support discrimination of rowing techniques. The stepwise procedure resulted in a model with five parameters ($F(5,474)=9.53$, $p<0.001$) explaining 9% of the variance of *hits* (true/false): both technical parameters (t_*grip* $p<0.001$, t_*grip/footrest* $p<0.05$) both amplitudes (grip $p<0.001$, seat $p<0.01$) and grip force maximum ($p<0.001$), but not foot rest force (both $p>0.05$). A stepwise approach with the dependent variable "rejections (true/false)" resulted in a much lower correlation of R-squared<0.006 ($F(1,1432)=9.50$, $p<0.001$), with significant contributions only of t_*grip/footrest*.

4. DISCUSSION

The purpose of the present study was to investigate perceptual effects of a complex movement sonification. Subjects heard movement sonifications of two consecutive rowers and had to estimate velocity-differences of their virtual boats. On a basic level this sonification provides information about two kinematic (grip and seat amplitude) and two dynamic parameters (grip and foot rest force) – parameters directly measured and mapped onto sound. Considering the continuous course of the parameters this sonification even provides information about temporal, biomechanical or technical parameters: Repeating pitch sequences provide information about rowing frequency; the time course of grip force informs about mechanical power; the time of a certain event in relation to other events reflects an individual technical pattern. Correlations between single parameters and the results of the perceptual task would suggest that higher percepts emerge from this sonification of the execution of own or foreign movements.

An interesting finding is that perceptual results were related to complex movement parameters. Variance of perceptual estimations was explained up to 80% by the parameter virtual boat velocity. Cohen [20] labeled correlations as large, as far as they explained more than 25% of variance. The much larger value of the present results therefore strongly suggests that this sonification has a large perceptual effect. Virtual boat velocity had not been mapped onto sound directly and therefore had to be derived on the basis of other parameters. Equation (1) points out that those parameters are related to displacements, time and forces, and correlation analysis show that the sonified parameters do not explain perceptual effects alone. Therefore it can be suggested that percepts emerged from combinations of those factors.

Coefficients of determination were in sum much larger than one and thus argue for a redundancy of information carried by the four sonified parameters. Further experiments might be necessary to reduce this redundancy or to identify the significant information content. But in contrast to this cognitive interest, the applicability of the sonification in training might profit from this redundancy: it gives the opportunity to chose among several parameters and to get sufficient results independent of the choice. The choice itself might depend on several factors as for example individual preferences, expertise, cognitive strategies or attentional focus. Therefore this redundancy could be of interest for experts, but first and foremost for non-experts as they have not learned to detect the most relevant movement parameters and to focus their attention on them.

The detection of the own movement yielded a *d'* of 0.47. Knoblich et al. [9] found in visual prediction tasks *d'*s of 0.34, 0.47 and 0.56, which is comparable to our detection task (task 2). But in contrast to our study those authors found in two experiments that subjects were just able to predict the outcome of self-induced movements, but not those from other persons. A possible explanation for the discrepancy: the prediction rate correlated negatively with the similarity of stimuli that had to be differentiated. When own movements and those of other persons were assimilated via instruction to perform in a defined way, predictions of other persons' movements became possible: Analysis of responses yielded a *d'* of 0.50, which was quite similar to the prediction rate of own movements. This finding sheds light on results of task 1: Movements on an indoor rower are constrained and limited to a few degrees of freedom. The standardization of rowing velocities adjusted and assimilated individual rowing techniques even more. Therefore, in line with

Knoblichs' interpretation, predictions of other movements should have been as good as predictions of own movements. This has exactly been found! Furthermore, when two different rowing techniques were presented within one trial (*unkown_different*), accuracy of predictions was significantly lower than when two similar techniques were presented (*unknown_same*).

Thus it can be concluded that own techniques and those of other persons can be well predicted by listening to movement sonification. This finding is supported by a final identification task, in which all rowers were asked to identify themselves and their named rowing partner after presenting two rowing cycles of five different persons: four athletes succeeded in identification of their *own* and three in the identification of the partner (*known*).

The results are compatible with the view that the own motor system is activated during the predictions of movement effects. The present study demonstrates large perceptual effects of movement sonification and most importantly, own techniques can be identified among others as good as in the visual modality. This suggests in line with former interpretations [8,9,11] that sonification can address motor representations. Latter conclusion is supported by a recent neurophysiological experiment: Schmitz et al. [21] could show that congruent movement sonification addresses the human object observation and mirror neuron system as well as key players of the motor loop. In that study congruent movement sonification was based on two kinematic parameters indicating that they carry sufficient information about the movement to address the mirror neuron and the motor system. Discrimination analysis of the present study supports this view. Two of five significant parameters provided information about spatial distances as in the above cited study. It is tempting to speculate that the technical parameters and information about grip force address motor representation too. But it could be criticized that hits were only predicted with a low to medium effect [19], even if the to-be-predicted own-effect is low. Nevertheless regression models could only predict decisions during presentation of own movements and not movements from other persons, indicating a linkage of those parameters to representations of own movements. Therefore a further study on these aspects including neurophysiological methods should be conducted.

4.1. Practical implications

The present study provided evidence for large perceptual effects of rowing sonifications and their potential to activate the own motor system just by listening to them. These and former findings [17,22] have practical implications. Vesper et al. [16] have shown that joint action – the coordination and synchronization of two or more people – succeeds if an athlete builds representations of his or her *own task* and the *movement goal*. Former studies have demonstrated that sonification can address both aspects: Novices learn more quickly and better to row when the rowing model and their own movements are sonified [17]. Thus they can build better representations of their *own task* than subjects that have to rely on visual perception or "natural" auditory information of the indoor rower. Another study chose a different approach as not movement techniques but movement effects were sonified: In a field-study Schaffert et al. [22] investigated whether the sonification of boat acceleration enhances boat velocity. Providing real-time feedback of boat velocity might help the athletes to build a common representation of the *goal* of their joint actions. By attending the common effect they might coordinate their movements in time yielding a common impulse. This hypotheses are supported by the finding of increased velocities [22].

The present results refer to a third mechanism for joint action addressed by sonification: building a representation of the *task of another person* [16]. Perceiving when and – most importantly - how other athletes move make their movement effects predictable as shown in task 1 of the present study. In consequence the synchronization of own and other movements could be even more effective. However, this is a hypothesis that will be investigated in further studies.

5. CONCLUSION

The results of the present study show that continuous sonification of two kinematic and two dynamic parameters provides enough information to predict the effects of complex movements and to identify the own technique among others. Further studies should investigate whether this kind of sonification can optimize synchronization of athletes.

6. ACKNOWLEDGMENT

Thanks are due to Markus Raab and Tanja Hohmann from the German Sport University Cologne and the University of Stuttgart for their helpful suggestions with respect to the design and analyses of the study, as well as to Klaus Scheerschmidt, rowing state coach, for his support and nomination of athletes.

7. REFERENCES

[1] A. O. Effenberg. "Movement Sonification: Effects on perception and action", *IEEE Multimedia*, vol. 12(2), pp. 53-59, 2005.

[2] M. Bangert, M., T. Peschel, G. Schlaug, M. Rotte, D. Drescher, H. Hinrichs, H. J. Heinze, and E. Altenmüller, "Shared networks for auditory and motor processing in professional pianists: Evidence from fMRI conjunction", *NeuroImage*, col. 30, pp. 917–926, 2006.

[3] L. Scheef, H. Boecker, M. Daamen, U. Fehse, M. W. Landsberg, D. O. Granath, H. Mechling, and A. O. Effenberg. "Multimodal motion processing in area V5/MT: Evidence from an artificial class of audio-visual events", *Brain Res.*, vol. 1252, pp. 94-104, 2009.

[4] G. Rizzolatti, L. Fogassi, and V. Gallese. "Neurophysiological mechanisms underlying the understanding and imitation of action", *Nature Rev. Neurosci.*, vol. 2(9), pp. 661-670, 2001.

[5] E. Kohler, C. Keysers, M. A. Umilta, L. Fogassi, V. Gallese, and G. Rizzolati. "Hearing sounds, understanding actions: action representation in mirror neurons", *Science*, vol. 297, pp. 846-848, 2002.

[6] A. Lahav, E. Saltzman, and G. Schlaug. "Action "Representation of Sound: Audiomotor Recognition

Network While Listening to Newly Acquired Actions", *J. Neurosci.*, vol. 27(2), pp. 308-314, 2007.

[7] S. M. Aglioti, P. Cesari, M. Romani and C. Urgesi. Action anticipation and motor resonance in elite basektball players", *Nature Neurosci.*, vol. 11, pp. 1109-1116. 2008.

[8] G. Knoblich, R. Flach, "Predicting action effects: Interactions between perception and action", *Psycol Sci*, vol. 12, pp. 467-472, 2001.

[9] G. Knoblich, E. Seigerschmidt, R. Flach, and W. Prinz. "Authorship effects in the prediction of handwriting cycles: Evidence for action simulation during action perception", *Q. J. Exp. Psychol.*, vol. 55, pp. 1027-1046, 2002.

[10] J. E. Cutting, and L. T. Kozlowski, "Recognizing friends by their walk: Gait perception without familiarity cues", *Bull. Psychonomic Soc*., vol. 9, pp- 353-356, 1977.

[11] F. Loula, S. Prasad, K. Harber, and M. Shiffrar, "Recognizing people from their movement", *J Exp Psychol Hum. Percept. Perform.*, vol. 31(1), pp. 210-220, 2005.

[12] P. E. Keller, G. Knoblich, and B. H. Repp, „Pianists perform better, when they play with themselves: On the possible role of action simulation in synchronization", *Conscious. Cogn.*, vol. 6, pp. 102-111, 2007.

[13] P. M. Bayes, and D. M. Wolpert, "Actions and consequences in bimanual interaction are represented in different coordinate systems", *J. Neurosci.*, vol. 26(26), pp. 7121-7126, 2006.

[14] N. Sebanz, N., and G. Knoblich, "Prediction in joint action: What, when, and where", *Top. Cogn. Sci.*, vol. 1(2), pp. 353–367, 2009.

[15] A. M. Williams, and K. Davids, "Visual search strategy, selective attention, and expertise in soccer", *Res. Q. Excerc. Sport*, vol. 69, pp. 111-128, 1998.

[16] C. Vesper, S. Butterfill, G. Knoblich, and N. Sebanz, "A minimal architecture for joint action", *Neural Networks*, vol. 23, pp. 998-1003, 2003.

[17] A. O. Effenberg, U. Fehse, and A. Weber. "Movement sonification: Audiovisual benefits on motor learning", BIO Web Conference, 1, 00022, 1-5. DOI: dx.doi.org/10.1051/bioconf/20110100022, 2011.

[18] A. Becker, *Echtzeitverarbeitung dynamischer Bewegungs-daten mit Anwendungen in der Sonification,* Unpublished thesis, Rheinische Friedrich-Wilhelms-Universität Bonn, 1999.

[19] D. M. Green, and J. A. Swets J.A., *Signal Detection Theory and Psychophysics*, New York: Wiley.

[20] J. Cohen, *Statistical power analysis for the behavioral sciences,* New York, London: Academic Press, 1969.

[21] G. Schmitz, B. Mohammadi, A. Hammer, M. Heldmann, A. Samii, T. F. Münte, and A. O. Effenberg, "Observation of sonified movements engages a basal ganglia frontocortical network", *Hum. Brain Mapp.*, under review.

[22] N. Schaffert, K. Mattes, and A. O. Effenberg. "Listen to the boat motion: acoustic information for elite rowers", *Proc. of ISon 2010, 3rd Interactive Sonification Workshop* Stockholm,, Sweden, 2011, pp. 31-37.

HEARING NANO-STRUCTURES: A CASE STUDY IN TIMBRAL SONIFICATION

Margaret Schedel

Department of Music,
Stony Brook University,
Stony Brook, NY, USA
mschedel@notes.cc.sunysb.edu

Kevin G. Yager

Center for Functional Nanomaterials,
Brookhaven National Laboratory,
Upton, NY, USA
kyager@bnl.gov

ABSTRACT

We explore the sonification of x-ray scattering data, which are two-dimensional arrays of intensity whose meaning is obscure and non-intuitive. Direct mapping of the experimental data into sound is found to produce timbral sonifications that, while sacrificing conventional aesthetic appeal, provide a rich auditory landscape for exploration. We discuss the optimization of sonification variables, and speculate on potential real-world applications.

1. INTRODUCTION

Sonification of datasets is becoming more popular as an alternative modality for exploring, and understanding, datasets. Beyond the obvious implications for accessibility, sonification enables interested parties to interact with data more deeply; e.g. multi-modal data exploration leverages more of a person's sensory 'surface area.' This is especially relevant in light of the modern trends in data collection: datasets are growing ever-larger, and in many cases ever-more complex, esoteric, and non-intuitive. We elected to study sonification of x-ray scattering data, which are rather abstract datasets that even experts struggle to understand.

An x-ray scattering experiment consists of directing a highly collimated, monochromatic, beam of x-rays through a sample of interest. The incident x-ray wave scatters off of all the atoms and/or particles in the sample, and the interference of these secondary waves produces scattered rays at angle that are characteristic of the material's internal structure. [1] In a scattering experiment, the deflection of scattered rays is characterized by the so-called *momentum transfer vector*, usually denoted by q, which is computed from the measured scattering angle, 2θ, by:

$$q = \frac{4\pi}{\lambda}\sin\theta \tag{1}$$

where λ is the wavelength of the x-rays. The quantity q has units of 1/distance, and q-space is thus frequently called 'inverse space,' or 'reciprocal space.' This abstract space is in some sense the Fourier transform of the realspace density distribution in the sample. Mathematically:

$$s(\mathbf{q}) = \sum_n f_n e^{i\mathbf{q}\cdot\mathbf{r}} \tag{2}$$

$$f_n(\mathbf{q}) = \int \rho(\mathbf{r}) e^{i\mathbf{q}\cdot\mathbf{r}} \mathrm{d}V \tag{3}$$

The scattered intensity, $s(\mathbf{q})$, is computed by summing the scattering contributions from the n scattering entities in the material (e.g. each atom). The scattering contribution of each entity, f_n, is in turn computed by integrating its density distribution, $\rho(\mathbf{r})$, over all of real-space.

Conceptually, the scattering experiment encodes all the information about the sample's shape and internal structure, albeit in an opaque and non-intuitive way. Roughly, a scattering peak at a particular q (i.e. angle) implies a real-space repeating structure with a size-scale of:

$$d = \frac{2\pi}{q} \tag{4}$$

We note that the inverse nature of $2\pi/d$ means that a scattering peak at large angle corresponds to small real-space distances, whereas a peak at small angle corresponds to larger real-space distances. As the field of nanotechnology matures, x-ray scattering is emerging as a powerful tool to study new materials; however interpreting this data is difficult. Although scattering data is in essence a Fourier transform of the material's structure, an experiment only captures the amplitude of the scattered waves, and cannot record the phase information.

X-ray scattering datasets are normally visualized using two-dimensional false-color images (see Figure 1). These images are an extremely valuable tool for researchers, but have their limitations. Scattering data can have a very large dynamic range, which is difficult to represent in a single image. Here, sonification can help, since the human ear has a correspondingly large dynamic range. [2] Moreover, the Fourier transform nature of scattering data implies a natural match with audio data. In scattering experiments, a given feature (e.g. at q_0) will frequently have harmonics (at $2q_0$, $3q_0$, etc.). Interpreting this axis as frequency in a sonification would naturally generate audio overtones which the human auditory system is exceedingly well-equipped to detect: timbre. Timbre is difficult to define, but has been described as "that attribute of auditory sensation in terms of which a listener can judge that two sounds, similarly presented and having the same loudness and pitch, are different." [3]

In this paper, we explore sonification as a tool to provide scientists with an additional method to deeply explore scattering datasets. The abstract nature of the data makes this a challenging, but critical, problem. Moreover the quantity of such data generated is growing hugely with time: newer x-ray instruments are now being built with ever-greater flux, generating data at an ever-increasing speed. It is also worth noting that scattering experiments can also be performed with visible light, electron beams, and even neutron beams. Although we focus very specifically on x-ray scattering data in this paper, we view this as a case study for the general problem of extracting meaning from the highly abstract datasets that are common in the physical sciences. We show that timbral

sonification generated directly from the data through additive synthesis [4] can provide a useful instantiation of abstract data.

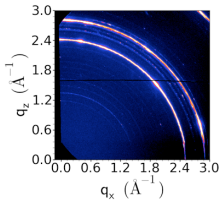


Figure 1: Example x-ray scattering data. The direct beam is incident near the lower-left corner of the image. The false-color image highlights certain features which arise from diffraction of x-rays from the sample's internal structure.

2. SONIFICATION

Over the last few decades, there have been a number of interesting cases of data sonification. Sonifications have been made of seismic data [5], ocean currents [6], and heart rates [7]. Despite these examples, sonification is a largely underutilized technique. Sonification provides a number of unique advantages: the human ear has a wide dynamic range across two variables: frequency and loudness; the human auditory system is finely tuned to detect subtle changes and extract signals from substantial noise; sonifications can be ambient, rather than requiring focused attention; and sonifications can be added to other forms of data exploration, creating more immersive multi-modal interactions.

Much of the existing work in sonification has involved conversion of time-series data. Such conversions are undoubtedly valuable, and are intuitive to understand, but this leaves aside the vast majority of datasets, where some non-temporal variable is of interest. In addition, recent sonifications have mapped the input data onto a tonal scale, or even used sampling or synthesis to reproduce notes from particular instruments. [8] These musical sonifications, like music itself, exploit pattern-seeking features of the human auditory system to create sounds that are crisp, distinct, recognizable, and typically pleasant. Although such realizations can be interesting, even beautiful, the musical nature frequently obscures the underlying patterns in the data. Herein we advocate for the more direct mapping between data space and sound. This necessarily leads to more complex, even cacophonous, sonifications; however such a mapping is relatively unbiased and preserves the majority of the information content. One can crudely identify a tradeoff between aesthetics and information content. Our sonification method uses pitch and loudness only to inform the additive synthesis; the main auditory channel is timbre.

We reformulate the two-dimensional scattering image into a (q, angle) array, where 'angle' is the arc angle with respect to the vertical axis of the image. In so doing, rings of scattering (which have a constant q-distance from the incident x-ray beam) are turned into straight horizontal lines in the $I(q,\text{angle})$ matrix. Doing so also highlights any variation in the ring intensity, which corresponds to spatial orientation of the structures in the sample. The intensity matrix has no time variable; we introduce time by in effect sweeping through the experimental data. In particular, the $I(q,\text{angle})$ matrix is directly converted into an $I(f,t)$ matrix, where f is frequency and t is time. This matrix is simply a spectrogram, or sonogram, which can of course be converted into a sound waveform through additive synthesis. For a sampling rate f_s:

$$A(t) = \sum_{n=1}^{N} I_n(t) \sin\left(\frac{2\pi f_n}{f_s} t\right) \quad (5)$$

here $A(t)$ is the instantaneous amplitude of the output waveform, and the $I(f,t)$ is discredited into $I_n(t)$ by splitting the frequency range into N bins. Thus the scattering data (the $I(q,\text{angle})$ matrix) is mapped directly into the amplitudes of the sine wave components of the sound. This synthesis inherently creates timbre-based (as opposed to tonal) sounds.

We wrote a simple program, using the Python programming language, which directly performs the computation in equation (5), and outputs the resultant waveform into a sound file. We note that this brute-force computation of the waveform is not necessarily the most computationally efficient, or elegant, means of performing additive synthesis (e.g. an appropriate FFT could be used). However we elected to use this method in order to provide flexibility in terms of redefining the mapping between the input data and the output waveform.

The mapping of q into frequency is extremely natural. As already described, both q and f are in some sense the variables along which a Fourier transform is taken. Both exhibit overtones and other natural relationships. The selected mapping is essentially taking the spatial modes (c.f. equations (2) and (3)) and mapping those into frequency modes. Although the one-to-one mapping between the $I(q,\text{angle})$ array and $I(f,t)$ array is information-preserving, and relatively natural, we must make a number of choices about what ranges to specifically map between.

3. PARAMETER OPTIMIZATION

In producing audio files from the two-dimensional data matrices, we must make a number of decisions about both audio encoding, and the range of the mapping (e.g. how to scale between angle and time). A sampling rate of f_s = 44.1 kHz (CD audio quality) was selected to provide sufficient quality for the detailed structures in the scattering data. Similarly, a 32-bit intensity encoding was used to allow for the large dynamic range of scattering datasets. As mentioned, there is a natural relationship between q and f. We align $q = 0$ with $f = 0$ so that any harmonics (or other natural progressions) in the scattering data are automatically converted into harmonics in the sound output. Scattering images are typically visualized using a false color map applied through a logarithmic scale, the human auditory system makes this unnecessary for sonification.

Further parameters were optimized by testing a variety of values. For this testing we used scattering data from a polymer solar cell material confined in a nanoscale grating (see Figure 2). Physically, this sample has an oriented morphology; this translates to a scattering ring whose intensity varies along the arc. This, in turn, translates into time variation of the sonification.

The mapping along the frequency axis, which encodes the q-values, is necessarily arbitrary. Although there is a natural reason to align the origins of q and f space, there is no

clear correspondence between inverse-distance and inverse-time units. The maximum frequency for human hearing is ~20 kHz. However this choice of frequency maximum was found to generate sounds with too many piercing components. Selecting too low a value for the frequency ceiling resulted in deep and rumbling sounds which essentially washed out all the structure in the scattering data. We found that an upper bound of ~5 kHz in frequency resulted in sonifications that were rich and preserved important data features, without leading to ear fatigue.

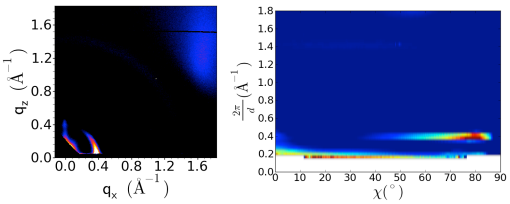


Figure 2: X-ray scattering data in false color on the left. The image on the right remaps the scattering data into an $I(q,\text{angle})$ matrix.

The partitioning of the frequency axis into N bins has a substantial effect on the quality and character of the final sound. An extremely low value (e.g. 10 bins per Å^{-1}), not surprisingly, over-smoothed the data and resulted in a loss of data. However, extremely fine partitioning (e.g. 1000 bins per Å^{-1}) introduced drastic beating artifacts into the sound. Essentially, by having more frequency resolution than actually warranted by the data's q-resolution, we introduce step-edges in the frequency envelope. The optimized value (50 bins per Å^{-1} for the test dataset) reproduces the spacing in the original data.

The construction of the $I(f,t)$ matrix also requires an arbitrary choice about temporal discretization. Note that this binning width is not the same as the sampling frequency, f_s. Whereas f_s describes the sampling rate used in the additive synthesis (the construction of the output waveform), the temporal binning describes the partitioning of the $I(f,t)$ matrix used to compute the amplitude values for the synthesis. The temporal resolution here is limited by the original dataset. As expected, using low temporal resolution (10 bins per second) smoothed over features in the data, effectively throwing away data. Higher data rates of course cure this defect. However, there is no advantage to increasing the time partitioning beyond that dictated by the initial data. We found that 50 and 1,000 bins per second were found to be essentially identical. We selected 150 bins per second as the optimal value, allowing a healthy safety margin. We improved the sound substantially by interpolating between the data points along the time axis. Doing so avoids sudden changes which introduce sharp popping artifacts into the sound, which hinders comprehension (not to mention damaging speakers).

The length of the sound has a strong effect on the listener's ability to discern structure. Sounds that are too short are difficult to parse. Stretching the sound helps reveal certain details, but inherently makes changes more gradual and difficult to notice. We found that sounds less than 1 second were too fast to be of any use. Sounds on the order of 1-2 seconds could potentially be useful for quick comparisons and identifications, but were still too fast to truly notice signal variations. At 3.5 seconds, sounds, and trends within those sounds, were discernible. Stretching sounds beyond ~10 seconds made it harder to track feature changes.

The above parameter optimization confirms certain limits of the sonification process, but is in some sense idiosyncratic to the datasets chosen. Ideally, all of these variables would be quickly and easily tunable by the user, allowing them to explore datasets in different ways. Looking forward, we envision a software interface that allows the user to select subsets of the scattering data to sonify, and allows the mapping ranges themselves to be easily modified.

4. VARIANTS

In the foregoing, we have attempted to motivate the use of the most direct, perhaps most naïve, mapping between the input data and the final waveform. We also explored a variety of alternative mapping strategies. Imposing additional mapping rules can be a powerful way to highlight certain features of datasets, and this is a valuable way to explore data through sound. We considered the following alternate mapping of intensity to waveform amplitude:

$$A(t) = \sum_{n=1}^{N} \sin\left(\frac{2\pi f_n}{f_s} I_n(t)\right) \qquad (6)$$

Here, rather than the intensities modulating the amplitude of the sine waves, they modulate the frequencies of these waves. By using the data matrix to modulate frequency, rather than amplitude, the character of the sound changes substantially. Changes in intensity become very strongly highlighted, as they produce noises that vary in pitch. These chirps or 'boomerang' sounds are distinctive and can be useful for uncovering subtle intensity changes, or small peaks, that might otherwise go unnoticed.

For many samples of interest in x-ray scattering, there is no preferred orientation of the material. Experimentalists typically convert these two-dimensional datasets into one-dimensional curves by averaging overall all possible angles in the image. Sonifying the original two-dimensional data using the approaches described above would result in a sound that does not vary with time. One obvious alternative mapping that we explored is to simply sweep time through the horizontal axis (q), and use the intensity to modulate the amplitude of a single tone at frequency f:

$$A(t) = I(t)\sin\left(\frac{2\pi f}{f_s} t\right) \qquad (6)$$

Although simplistic, this mapping can be useful. In particular, the existence of equally-spaced peaks in scattering data yields a metered oscillation in the sound. Moreover, subtle deviations of peak positions could be picked up by the listener, as hearing is able to discriminate small timing differences. As with the two-dimensional data, we can use the intensity data to instead modulate the frequency of the sound:

$$A(t) = \sin\left(\frac{2\pi f}{f_s} I(t)\right) \qquad (7)$$

Here again, we discover that by modulating frequency, rather than amplitude, sudden changes in intensity in the data become highlighted by sweeping changes in frequency. Details of peak positions and heights are sacrificed, but extremely weak peaks now become readily apparent. This points again to the need in sonification for user-adjustability.

5. APPLICATIONS

Scientists studying x-ray scattering have already developed a sophisticated toolbox of visualization techniques to explore data, and theoretical models to explain, quantify, and fit their data. It is thus natural to ask whether sonification can bring any new insight to the task of understanding these abstract datasets. We envision a variety of ways in which sonification could elucidate experiments. Consider the data shown in Figure 3, for four different kinds of samples. The false-color images are all quite distinct; and indeed the corresponding sounds are all unique and extremely distinct: the first image has many striations which leads to a number of fairly distinct tones persisting in time. The second image is a 'misaligned' sample; the corresponding sonification is dominated by blips and cracks that sound distinctly like artifacts. The third example is a composite of nanotubes dispersed in an elastic polymer. The scattering image has diffuse intensity throughout, due to the disordered arrangement in the sample; this can be heard as a hazy, wind-like sound permeating the sonification. The final example is a nano-scale grating. Here, the extremely regular and precise structure results in many distinct streaks in the false-color image. These streaks create periodic rhythms in the sonification.

One notable advantage of sonification over careful visual inspection is that the former can be done ambiently. Modern scientific instruments are becoming increasingly automated, to handle the growing scale of scientific discovery. Sonification provides the opportunity for the experimenter to work on other tasks, while listening, in the background, to automated data collection. Any sudden changes in the incoming data, or surprising samples, will immediately be noticed and can be explored in greater detail. Consider for instance the 'misaligned' sample; the sonification is distinct and the experimenter would immediately know that something was wrong with the instrument.

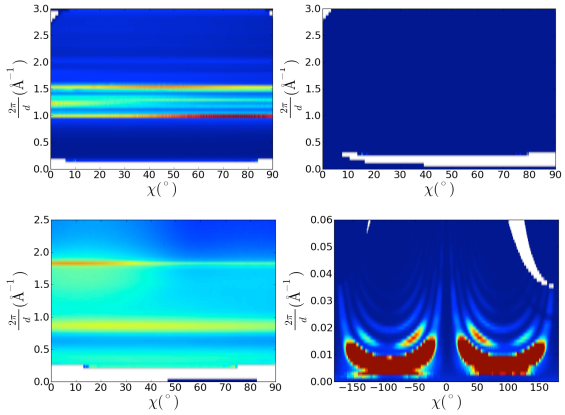


Figure 3: Examples of the variety of data one can obtain from x-ray scattering. From top to bottom the sample are: a semi-crystalline commercial plastic; a 'misaligned' sample (where the beam missed the sample); a composite of carbon nanotubes in a matrix of elastic polymer; and an empty nano-scale grating.

With some effort and training, it is also likely that an experimenter could learn to differentiate between all the unique features in the sound, and could pull out interest trends and features that they had ignored in a visual analysis. It is clear, however, that what is lacking are fast and easy-to-use software tools to enable users to quickly explore different mappings and different datasets.

6. CONCLUSION

We have presented a case study of sonifying x-ray scattering data. Direct mapping of the two-dimensional intensity values of a scattering dataset into the two-dimensional matrix of a sonogram is a natural and information-preserving operation that creates rich sounds. Our work supports the notion that many problems in understanding rather abstract scientific datasets can be ameliorated by adding the auditory modality of sonification. We further emphasize that sonification need not be limited to time-series data: any data matrix is amenable.

Timbral sonification is less obviously aesthetic, than tonal sonification, which generate melody, harmony, or rhythm. However these musical sonifications necessarily sacrifice information content for beauty. Timbral sonification is useful because the entire dataset is represented. Non-musicians can understand the data through the overall color of the sound; audio experts can extract more detailed insight by studying all the features of the sound.

7. ACKNOWLEDGMENT

X-ray scattering experiments were carried out on the X9 beamline, which is managed by the National Synchrotron Light Source and the Center for Functional Nanomaterials, Brookhaven National Laboratory, which are supported by the U.S. Department of Energy, Office of Basic Energy Sciences, under Contract No. DE-AC02-98CH10886. We thank Danvers Johnston for providing the test sample for the scattering experiment.

8. REFERENCES

[1] B.E. Warren, *X-Ray Diffraction*, New York, USA: Dover Publications, 1990.

[2] T. Fitch & G. Kramer, G. *Sonifying the body electric: Superiority of an auditory over a visual display in a complex, multi-variate system.* In G. Kramer (ed.), Auditory display: Sonification, audification and auditory interfaces. Proceedings of the First International Conference on Auditory Display (ICAD) 1992, 307-326. Reading, MA: Addison-Wesley, 1994.

[3] J. Neuhoff, "Perception, Cognition and Action in Auditory Displays." In T. Hermann, A. Hunt, J. Neuhoff. (ed.), *The Sonification Handbook*. Berlin: Logos Publishing House, 2011.

[4] C. Roads, *The Computer Music Tutorial,* Cambridge, MA: MIT Press, 1996.

[5] J. Acoust. *Seismometer Sounds.* Soc. Am. Volume 33, Issue 7, pp. 909-916, 1961.

[6] B. Sturm. "Pulse of an Ocean: Sonification of Ocean Buoy Data." *Leonardo*, Vol. 38 Issue 2, pp. 143-149, 2005.

[7] M. Ballora et al., "Heart Rate Sonification: A New Approach to Medical Diagnosis," *Leonardo* Vol. 37, No. 1, pp. 41–46, 2004.

[8] J. Kreidler *Charts Music – Songsmith fed with Stock Charts.* http://www.youtube.com/watch?v=2-BZfFakpzc&feature=player_embedded, 2009.

SONIFICATION OF A REAL-TIME PHYSICS SIMULATION WITHIN A VIRTUAL ENVIRONMENT

Rhys Perkins

Anglia Ruskin University,
Department of Music and Performing Arts,
Cambridge, CB1 1PT, UK
rhysperkins@rhysperkins.com

ABSTRACT

There has been an increasing amount of research utilising 3D virtual environments as a core component of interactive sonifications. While showing considerable potential for their ability in producing both real-time visualisation and sound, they often come with constraints as a result of their design decision processes. This paper presents developments of a prototype that has arisen out of my attempts to address some of the issues involved in bringing sonification to a wider audience through a universal metaphor. These new additions allow for an intuitive, elementary introduction into to the world of auditory display, while providing a more flexible and immersive environment for composition and sound design.

1. INTRODUCTION

The emergence of improving technology has provided an opportunity to overcome the limitations of previous work [1] where computational requirements would produce significant latency between the audio and the visual, inhibiting real-time interaction. Dedicated hardware, such as the graphics processing unit (GPU), allow for the sharing of resources and have recently shifted their focus towards more general purpose computing [2] including accelerating physics simulations. Freely distributed 3D physics engines such as Bullet[1] and Open Dynamics Engine,[2] common middleware solutions for modern game engines, are readily available thanks in part to an increasing demand for realism amongst gamers. Their cross-platform approach sees widespread use of long established solutions such as the desktop computer but also extends to mobile devices. As a result this presents an opportunity to revisit previous compositional techniques [1] and reach a wider audience in the process.

Nguyen's approach to TONAL DisCo [3] acknowledges the benefits of using a game engine when combining dynamic visuals with audio but chooses to use a pre-processed sample library over real-time sound generation. The prototype laid out in this paper addresses that limitation by utilising a messaging system that provides a link between the visual and the sound synthesis. Pre-processed sample libraries alongside dynamic visuals have started to emerge in commercial software with applications like PhysSynth[3] for the iPad. However, I would consider this application to be compromised by only simulating basic particles with simple collision detection and response, comprising interactions between points and segments and restricting the process to two dimensions. By sonifying 3D physics engines we can surpass these restrictions to cover more interesting and complex interactions, communicating and exploring a wealth of new dynamic ideas within an auditory display whose interface with the user resembles the physical world.

Sturm [1] laid out some of the benefits that a sonification of particle physics simulation would bring to science, including new ways of understanding physical phenomena, and refers to several artistic merits such as bending of those scientific laws to suit a composer's taste. Pedagogical advantages that Western music composition students might gain from using an audiovisual simulation are also discussed. In particular, he suggests there is an advantage to combining the audio and visual modalities in order to present musical ideas, likening it to listening to a piece whilst reading a score, rather than partaking in one or the other activity separately. Metaphorical correlations between particle physics and sound synthesis have been explored [1][4], serving as a means of providing a bridge between the two aforementioned fields whilst highlighting cultural differences and the problems that might arise from them. For instance Sturm [1] states that composers must possess skills in physics to begin with and only the audience need not be versed. The system described in this paper was designed to cater for all levels, from the well-versed user who comprehends and wishes to explore and extend the open-source scientific algorithms to those that would like to immediately compose and play.

RedUniverse provided a toolkit for sonifications of dynamic systems [5] with the aim of producing a playground for compositional ideas. These systems were also limited to two dimensions and lacked accessible interactivity, requiring a good knowledge of the SuperCollider programming language in order to take full advantage of their potential customisation. What I present here will allow for immediate use and configuration via standard input, such as a keyboard and mouse, but will also cater for further inputs using standard protocols such as MIDI and OSC.

Interactive compositional tool, VR-RoBoser [6] makes a case against predetermined, repetitive soundscapes in a virtual environment by using a context dependent sonification. They present the idea of a user-controlled or autonomous avatar that continuously reacts to its unchanging surroundings in order to overcome this issue. I would argue that the dynamic nature of physics would help create a less static environment. Continuous user interaction would stimulate audible results as the simulated objects react accordingly. Automated movement can be

[1]http://bulletphysics.org/wordpress/
[2]http://www.ode.org/
[3]http://www.physynth.com/

accomplished through inter-object logic allowing the user to pay attention to other aspects of this system, such as the proposed camera control.

In previous work [7] I laid out the design process behind a musical tool employing a 3D space that could be populated with audiovisual objects acting under the laws of physics. Each object exposed fundamental data dimensions that could then be mapped to sound dimensions via the OSC protocol, providing an audible insight into their behaviour within the current environment. A modular approach to the object design meant that a user would construct logic through object interaction and association without the need for coding keywords, operators and the understanding of basic programming paradigms.

The tool comprised an environment where polygonal models, visually representing the underlying simulation data, were introduced alongside a graphical user interface (GUI) that offered the opportunity to control fundamental properties, sonify and compose in real-time. Since designing the initial prototype I have found the need for improvements in a few key areas. This paper will discuss some of the fields I have considered when refining my approach to the sonification of an interactive physics engine. In the next section I will explore the idea of human interaction and how the process came to shape the design of the basic objects. The objects are then discussed along with some thoughts on their potential behaviour and how this can affect the overall output. This is followed by the mapping section which explains why I believe that the same simulated objects are inherently easier to map due to their physical grounding, along with several theories that underpin the conception of the mapping function tool. The camera section describes how the user will view and traverse the environment, how it accentuates the user's experience and why they should be presented with the option for both automatic and manual control configurations. This then leads into the messaging system where the aim is to make the same experience more flexible and personal. The paper concludes by stating some of the advantages the proposed system has over another comparatively close environment [8] along with my thoughts about potential future work and emerging areas of interest.

2. KEY COMPONENTS

2.1. Human Interaction

Investigation [9] has shown that applying human interaction in real-world contexts to sonification can help improve interface design. In this paper, the researchers state that humans are adapted for interaction within their physical environment and making continuous use of all their senses. When we perform an action on an object we expect some kind of reaction and our perception of objects builds up over time through this interactive process. The objects found in this system adhere to the unchanging laws of physics that our neural hardware has been effectively programmed to deal with over many years of evolution. This enables the user to utilise their acquired skills in order to manipulate high-dimensional data via objects with familiar behaviour and response. The authors also argue that one of the main problems in the domain of data exploration is that the data often inhibits a high-dimensional data space that is different from the 3D space we are familiar with. In this prototype the simulation data emulates rigid bodies, and their behaviour in our natural environment, providing familiar grounds for both exploration and interaction.

Interaction with the objects has a direct effect on the procedurally generated simulation in a similar manner to model-based sonification [10] where the user supplies the initial excitation. By grabbing, moving and throwing objects it is feasible to perform a wide range of actions from striking, to more delicate procedures such as plucking; each of which results in changes to the data dimensions. This direct process introduces information manipulation to the average user at a more accessible level when compared to other similar applications [11] since no coding knowledge is required. For example, saved scenarios containing preconfigured entities can be loaded, ensuring that user interaction yields instant audible results.

If the user wishes to create and save their own scenario, or edit existing ones, some basic GUI control knowledge is required. The controls can be toggled at any time and aim to present a more traditional and precise means of modifying the attributes that influence each aspect of the prototype. Presentation of the data in this manner brings its own set of problems in that the potential to overwhelm the viewer with information is increased. When considering high-dimensional spaces one study [12] argues for a mental model simpler than brute-force awareness of every detail in order to avoid cognitive overload. The authors suggest that parameters should be cross-coupled so that the performer naturally thinks of certain parameters as varying together in predefined patterns. The high-dimensional data, encapsulated visually by each model, allows us to intrinsically understand how the parameters vary together. Throwing an object would imply a change in velocity that would be influenced by the mass of the object. Spherical objects are more naturally inclined to roll, providing smoother changes in angular velocity as opposed to the sudden, erratic changes of their square shaped equivalents.

Research into improving sonification tools [4][9] has questioned how information should be distributed to different modalities in order to maintain the best usability. As stated previously, our everyday interactions with physical objects providing a base level for our conceptual understanding of the data dimensions found within. With this in mind I highlighted what I believed were the important elements of the underlying data, choosing to expose those that had a direct impact on the representative model's behaviour. Given the longstanding synergy between humans and physical entities I would suggest that less mental bandwidth is required to comprehend the visual events. Instead, the attentive capacity of the user can focus on the audio, and its governing mapping process, encouraging sonic exploration and creativity.

2.2. Objects

Objects provide a modular approach to the way the user experiences the underlying data. Depending on the object's configurable physical parameters the program will automatically simulate subsequent interactions as the object reacts to its current environment and user intervention. However, it can be argued that there are parameters that have no direct effect on the simulation which are just as important for the user to exploit. These properties can enhance a user's experience, and encourage them to learn, by creating associations through further visual abstractions that can be audibly reinforced. As one example, the

object colour could be changed in order to present object information in a new manner. According to research in Gestalt laws of grouping [13] there is a stronger tendency to group local elements by common colour than by similarity of shape. This would imply that, in some cases, our brains are more receptive to the material that encompasses each shape, rather than the shape itself. Therefore, by involuntarily grouping similar coloured entities, the audience's attention would be drawn to a single contrastingly coloured object, perceiving it as being outside of the group. The performer could then take advantage of this visual phenomenon by using it to introduce a solo theme or demonstrating object-specific sonic behaviour.

In a typical physics simulation most objects will likely come to rest until excitation provides the impetus for movement. If frequent changes in data are desired then further logic can be introduced via object specific context menus (Figure 1). For example the user could define a point where rigid bodies can spawn at regular intervals. Each body created would have a lifespan where the associated object would be automatically removed from the scene after such time had elapsed. Inter-object logic can be extended further by defining connecting mechanical joints and introducing external forces such as gravitational fields. Automated mechanical contraptions would be a logical step in complexity, allowing for the creation of visual algorithms. With the basic building blocks, it should be possible to conceive and construct contraptions in the style of Heath Robinson[1] or Rube Goldberg[2], providing unfamiliarity through the extraordinary.

2.3. Mapping

In this prototype the parametric mapping process grants an insight into the composer's conceptual understanding of the data dimensions. It has been suggested that metaphors help create more intuitive mappings [14] and is well suited to parameter mapping sonification [4]. Whereas the universal laws of physics can represent a predictable visual behaviour by employing a metaphor that fits our everyday observations, the sound representation is more subjective. The mapping of the objects serves to reflect the experiences of the user, making it difficult to produce general metaphors that are valid in any context. What may be coherent and intuitive to one mindset could be judged differently by those from another cultural background. There have been attempts to create online databases [15][16] suggesting mappings based on experimental evidence although it is widely accepted that an affective mapping can't always be predicted [4][14]. A heuristic approach to this area should be adopted to allow for a compositional process that encourages experimentation in order to express creativity where the audience can reflect on the implications of a musician's cultural and physical experiences.

When interviewing scientists, Vogt and Holdrich [4] discovered that strong metaphors emerge from their professional experience. They found that more mapping associations were suggested for the well-known particles and fewer for the rarer proposing that perhaps this arose from fewer encounters, lack of interaction, and therefore less prominent in the mind. They also discovered that everyday properties such as mass were cited more often than abstract ones. This would imply that an object visually described through a recognisable metaphor, encapsulating everyday properties, can be easier to map.

Our experience with physical objects allows us to inherently determine complex data relations. Properties are implied by a rigid body's response to collisions with its surroundings. By referring to the visual behaviour during this event we perhaps reduce the need to refer to the linking of parameters for interpretation. This can be illustrated by focusing on two dimensions, such as mass and velocity, where one could map them to pitch and envelope time, respectively. We understand that an object of greater mass would provide an object of less mass with a higher velocity upon collision. A spectrum of sound can be obtained afterwards where we could assume that objects that have travelled further will differ in pitch, and duration, to those that travelled less distance over the same period of time.

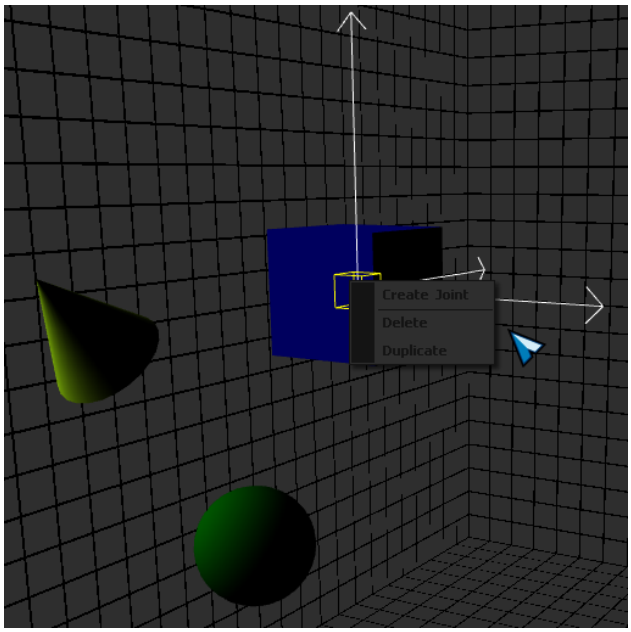


Figure 1: Three objects in the environment. The context menu is displayed over the selected object.

Understanding of the meaning in sonifications depends on the metaphors implied where the choices made during the process are crucial for how a design is understood by its listeners [17]. For instance, the coupling of coloured objects mentioned in section 2.2. Walker [18] states that in order to achieve an effective mapping choice, one must go beyond that of polarity and linear scaling functions while avoiding restrictions placed on the user through bad design [19]. The mapping window controls were devised to encourage flexibility by employing a messaging system, discussed later in this paper, to allow the user to map exposed parameters to potentially any input of a synthesiser. In conjunction with these GUI controls (Figure 2) I created a function editor that serves to display the relationship between the two dimensions. The editor itself contains two permanent breakpoints that define the input domain (x axis) and the output range (y axis). Further breakpoints can be added and removed in order to construct a bijective mapping curve or polyline. The curvature of the segments, found between each breakpoint, can also be configured in order to account for both linear and non-linear responses.

[1]http://heathrobinson.org/exhibition/index.htm
[2]http://www.rubegoldberg.com/

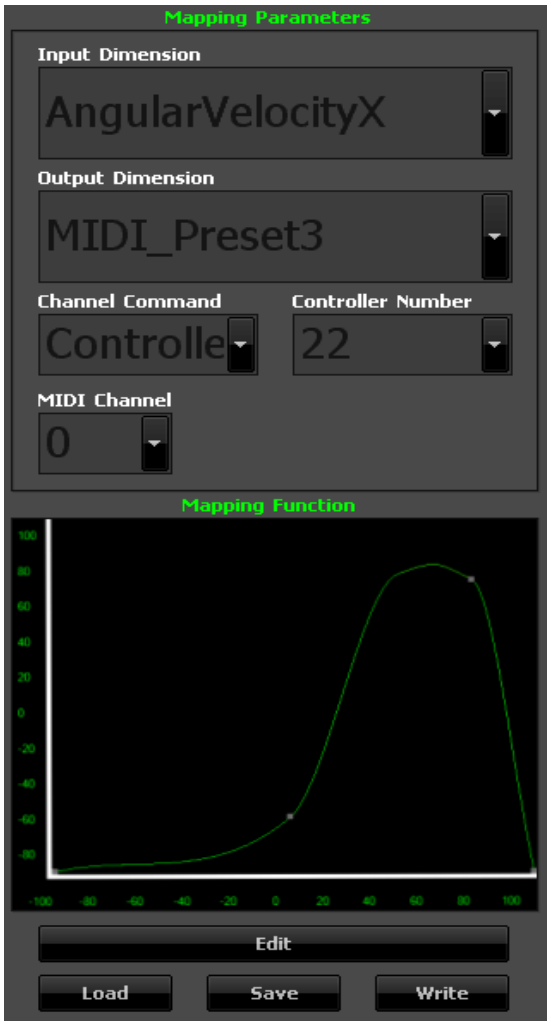


Figure 2: The mapping window along with the available controls for editing the relationship between the input and output dimensions.

2.4. Camera

Most software synthesisers use flat imagery in order to represent modifiable parameters making use of the two dimensions provided by a computer screen. Their interfaces often restrict the information to a window, presenting the intended audience with a multitude of controls that are difficult to decipher and engage with. The prototype's main interface attempts to address this problem, presenting the information in a more natural three dimensional form. This posed the question of how one might traverse the extra dimension in an immersive and inherent manner using the same display hardware.

Many first person perspective games attempt to provide immersive, high-level simulations of reality. In a typical setup, the player views, navigates and interacts with the game world by operating a camera. Recent studies have compared games in this category to sonification systems [20], where sounds are used to accentuate the player immersion by reacting to their behaviour or to provide sign posts for orienteering. I felt it was appropriate to adopt some of the ideas found in these proven systems by implementing a similar camera system for interactive traversal of the virtual environment.

Camera movement can be kinaesthetically controlled in real-time or automated along a user-defined pathway. Default manual control is that same as that of a typical PC first person shooter setup, utilising the W, A, S and D keys for camera translation, and a mouse for camera rotation. The camera rotation system required slightly different approaches to each mode of operation as manual rotation of a camera with six degrees of freedom would be disorientating with standard mouse control. If the camera pitch was allowed to be greater than 90 degrees in either direction, the mouse controls would be reversed along both the pitch and yaw axes. I therefore decided to emulate more natural head movement by restricting the camera pitch to a ±90 degree range in the same manner that a first person perspective camera does.

Node Property	**Description**
Position	Location of the node
Rotation/Orientation	Camera's orientation when reaching the node
Speed	The constant speed of the camera until the next node is reached *Time will be recalculated*
Time	The time (seconds) at which the camera arrives at the node *Speed will be recalculated*

Table 1: Configurable properties of an automated camera node.

Automated camera control frees the user from direct control giving them the opportunity to concentrate on other tasks, such as object interaction, and does not require the same restriction for rotation. Camera motion is defined by a series of nodes that comprise a Hermite spline-based path. The properties of each node (Table 1) allow for an increase in the accuracy and response time of the camera when compared to independent user control. Spherical linear interpolation is employed to ensure smooth changes in camera orientation when moving from one node to the next and prevents viewer disorientation. The timing of the camera can add to the overall sense of structure, guiding the viewer to focus on visual snapshots of the environment at designated points in time where precise values, for both speed and time, support various tempi.

By directing the camera, the user can create a sense of motion, guiding the audience through a visual soundscape. Choices made in constructing the camera's pathway become part of the creative process, enabling the viewer to observe through the cognitive lens of the composer. In this manner, attention can be drawn to specific areas of interest whilst providing an insight into the structures underlying the composition. Sturm [1] touched on this particular benefit of a camera system when he stated 'thus any sonification of a particle system is dependent on

the state of the observer; each observer with a unique position and/or velocity will hear the system in a different way – a truly relativistic idea.'

2.5. Messaging System

The messaging system sends and receives OSC and MIDI based messages providing the user with an opportunity to customise the data flow both in and out of the software. Whereas traditional human interface devices, such as a mouse and keyboard, can be used without the need for this system, these two protocols provide a widely accepted standard for interface control, expanding upon the breadth of possible controllers and the levels of immersion they provide. For example, tactile feedback can be introduced by utilising this system. Specific data, sent when an object is grabbed, can be interpreted by the current controller in order to produce corresponding actions such as vibrating.

With the mapping system controlling the outgoing data, I separate the audio processing from the software. Implementing these protocols grants access for communication across a network, extending the reach to computers operating on different platforms, such as OS X or Linux, or mobile platforms such as Android or iOS. This would provide the opportunity to communicate with a vast range of audio applications and synthesis tools, exploiting their existing timbres and increasing potential sonic diversity when compared to an inbuilt synthesis engine.

I have produced a video [21] that demonstrates an example of my prototype connected to sixteen instances of the Alchemy VST plug-in, utilising all available MIDI channels on a loopback network. The table shown below the video contains the mapping details for each object and was automatically generated by the prototype.

3. CONCLUSION

This work sets out to add to the range of tools for experimentation and interaction of a data set using the combination of the visual and aural modalities. It intends to expand on the experience found in Versum [8] by incorporating objects that dynamically respond to real-time interaction, with a unique collision response determined by their configurable properties. Whereas the entities found in the aforementioned software lend themselves to more ambient sounds due to their continuous playback, the objects here also allow for more dynamic sound with full control of ADSR envelopes and sonic response. Furthermore, the flexible mapping system does not constrain the amplitude of objects based solely on their distance from the camera as other relationships can be explored via the messaging system.

Whilst much time has been spent creating the basic tools for both music creation and sonification, I feel that future work should be focused on musicality. Sonifying data into a systematic musical structure to understand patterns and trends in a more traditional sense would have the benefit of making the tools more widely understood by the potential audience.

Nguyen discusses an approach to musicality in the mapping process where he decides to lose resolution of the data in favour for an increase in musicality [3]. He argues that the use of musical structure in sonification has the potential to communicate compound relationships with an increase in clarity that might not be apparent with high resolution data. To integrate this I would suggest changes to the function editor that would accommodate user-defined bands of any width. These regions would then be displayed on top of the mapping transfer function (Figure 3), allowing the user to conceive the varying resolutions. For example, on the y axis, the output dimension of pitch can be constrained to a musical scale (Lydian, Chromatic, etc.) using standard frequency tuning. This idea can be extended into other areas such as rhythm where the triggering of sound can be quantised to match common subdivisions of a bar based on the global tempo assigned. The effect should be subtle as to not lose perceived concurrency between the audio and the visual events.

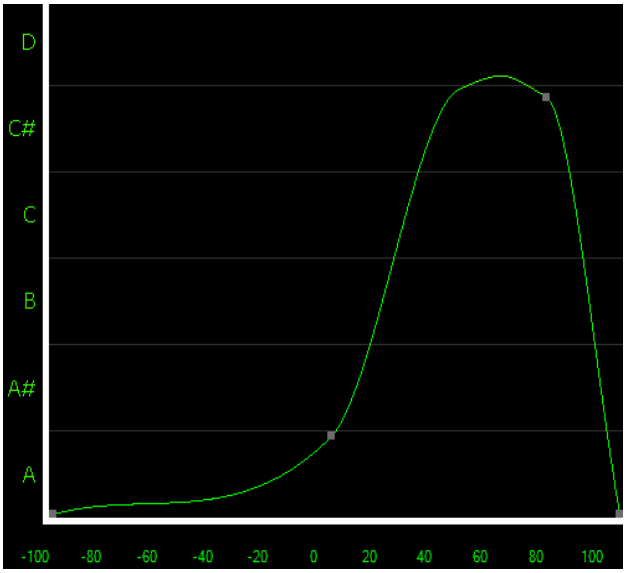


Figure 3: A function with granular regions denoted by the horizontal lines.

Recent developments in human computer interfaces, including the 'See Through 3D Desktop'[1] and the HoloDesk,[2] present more direct ways of interacting with virtual 3D objects. Model-based sonification has been shown to be intuitive by taking important dimensions of sound semantics into account and grounding them in physical sound generating processes in a natural and user-transparent way [22]. By combining projections of the simulated objects with interfaces that emulate a more innate way of interaction, we can extend the model-based method beyond its inherent physical constraints. This would benefit the interface building process as a variety of deformable, polygonal objects could be designed, created and saved in a portable format. It would also encourage the creation of more abstract and imaginative virtual controller shapes whose physical counterparts would be difficult or impossible to implement.

4. REFERENCES

[1] B. L. Sturm, "Synthesis and Algorithmic Composition Techniques Derived from Particle Physics," in *Proceedings*

[1]http://leejinha.com/See-Through-3D-Desktop
[2]http://research.microsoft.com/apps/video/dl.aspx?id=154571

of the Eighth Biennial Symposium on Technology and the Arts, Connecticut College, New London, Connecticut, 2001.

[2] M. Harris. "GPGPU: General-purpose computation on GPUs," in *Game Developers Conference*, San Francisco, California, 2005.

[3] V. Nguyen, "Tonal DisCo: Dissonance and Consonance in a Gaming Engine," in *Proceedings of the International Conference on Auditory Display*, Budapest, Hungary, 2011.

[4] K. Vogt and R. Holdrich, "A Metaphoric Sonification Method – Towards the Acoustic Standard Model of Particle Physics," in *Proceedings of the International Conference on Auditory Display*, Washington, USA, 2010.

[5] M. d'Inverno, F. Olofsson, "RedUniverse – a simple toolkit," in *Live Algorithms for Music Conference*, London, UK, 2006.

[6] S. Le Groux, J. Manzolli, P. Verschure, "VR-RoBoser: Real-Time Adaptive Sonification of Virtual Environments Based on Avatar Behavior," in *Proceedings of the Conference on New Interfaces for Musical Expression,* New York, USA, 2007.

[7] R. Perkins, "Mapping 3D Objects To Synthesised Sound Using A Simulated Physics System," in *Proceedings of the International Computer Music Conference*, Huddersfield, UK, 2011.

[8] T. Barri, "Versum: Audiovisual Composing in 3D," in *Proceedings of the International Conference on Auditory Display*, Copenhagen, Denmark, 2009.

[9] A. Hunt and T. Hermann, "The Importance of Interaction in Sonification," in *Proceedings of the International Conference on Auditory Display,* Sydney, Australia, 2004.

[10] T. Hermann, "Sonification for Exploratory Data Analysis," PhD thesis, Bielefeld University, Bielefeld, 2002.

[11] I. Bukvic and K. Ji-Sun, "μ Max-Unity3D interoperability toolkit," in *Proceedings of the International Computer Music Conference*, Montreal, Canada, 2009.

[12] G. Garnett and C. Goudeseune, "Performance Factors in Control of High-Dimensional Spaces," in *Proceedings of the International Computer Music Conference*, San Francisco, 1999.

[13] P. T. Quinlan and R. N. Wilton, "Grouping by proximity or similarity? Competition between the Gestalt principles in vision," in *Perception 27*, 1998, pp. 417–430.

[14] B. Walker and G. Kramer, "Mappings and Metaphors in Auditory Displays: An Experimental Assessment," in *ACM Transactions on Applied Perception*, vol. 2, issue 4, 2005.

[15] S. Barrass, "EarBenders: Using Stories About Listening to Design Auditory Interfaces," in *Proceedings of the First Asia-Pacific Conference on Human Computer Interaction*, Information Technology Institute, Singapore, 1996.

[16] B. Walker, and D. Lane, "Sonification Mappings Database on the Web," in *Proceedings of the International Conference on Auditory Display*, Espoo, Finland, 2001.

[17] A. de Campo, J. Rohrhuber, T. Bovermann and C. Frauenberger, "Sonification and Auditory Display in SuperCollider," in *The SuperCollider Book,* S. Wilson, D. Cottle, and N. Collins, eds., 2011, Cambridge, UK: The MIT Press, pp. 381-408.

[18] B. Walker, and D. Lane, "Psychophysical Scaling of Sonification Mappings: A Comparison of Visually Impaired and Sighted Listeners," in *Proceedings of the International Conference on Auditory Display*, Espoo, Finland, 2001.

[19] G. Kramer, et al. "Sonification Report: Status of the Field and Research Agenda," in *Report prepared for the National Science Foundation by members of the International Community for Auditory Display*, Santa Fe, NM, 1999.

[20] M. Grimshaw, "Sound and Immersion in the First-Person Shooter," in *Proceedings of the International Conference on Computer Games*, La Rochelle, France, 2007.

[21] R. Perkins, "Video demonstrating MIDI out capabilities of the prototype," Available at www.rhysperkins.com/Sonification/MIDI.htm, Accessed 8 February 2012.

[22] T. Hermann, H. Ritter, "Sound and Meaning in Auditory Data Display," in *Proceedings of the IEEE*, vol. 92, issue 4, 2004.

CircoSonic: A SONIFICATION OF CIRCOS, A CIRCULAR GRAPH OF TABLE DATA

Vinh Xuan Nguyen

University of New South Wales,
Faculty of Built Environment,
2022, Sydney, Australia
vinh.x.nguyen@unsw.edu.au

ABSTRACT

This paper presents, applies and evaluates "CircoSonic," an interactive sonification of "Circos." It outlines the development of modifying a gaming engine to replicate Circos, a circular graph for comparing pair wise relationships in a 2D data table, with the added capabilities of sonification through interaction.

The developed prototype is applied to a static dataset and evaluated using an insight based methodology. The evaluation uses a muted version of CircoSonic to establish a comparison between visualizations, from which a comparison between visualization and sonification can be extrapolated.

The results demonstrate that with a static dataset, CircoSonic with sound consistently outperforms CircoSonic without sound, and Circos, despite being solely visual, outperforms both versions of CircoSonic. The conclusion is that the visualization component of CircoSonic can be significantly improved and that a move from static to dynamic data may display different results. The investigation of novel visualizations from the perspective of auditory displays needs to be extended to include those which deal with multivariate and dynamic datasets whilst still offering a broader application to diverse data domains.

1. INTRODUCTION

The field of auditory displays, through the course of its development, has used the field of information visualization as a point of comparison, often referring to Tufte [1-11]. Auditory displays aim to accomplish similar goals as visual displays through the alternative, but often complementary communicative medium of sound which affords advantages that are difficult or impossible with vision.

The most common visualizations, such as line graphs, bar graphs and pie charts, have been investigated by researchers of auditory displays and sonification. However, the emergence of novel scientific visualizations necessitates an investigation into less common but more potent visualizations. One emerging visualization exemplary is "Circos," a circular graph for comparing pair wise relationships in a 2D data table [12]. It has been used by "mainstream periodicals and newspapers" [13-16] to communicate to a general audience. Circos (see Fig. 1) was developed for the comparative genomics field as a visualization tool, but is also applicable to other data fields where identifying relationships and the nature of those relationships are of interest. The Circos graph in Fig. 1 (Created with Table Viewer: http://mkweb.bcgsc.ca/tableviewer) is a visualization of the data in Table 1. The graph shows a relationship between

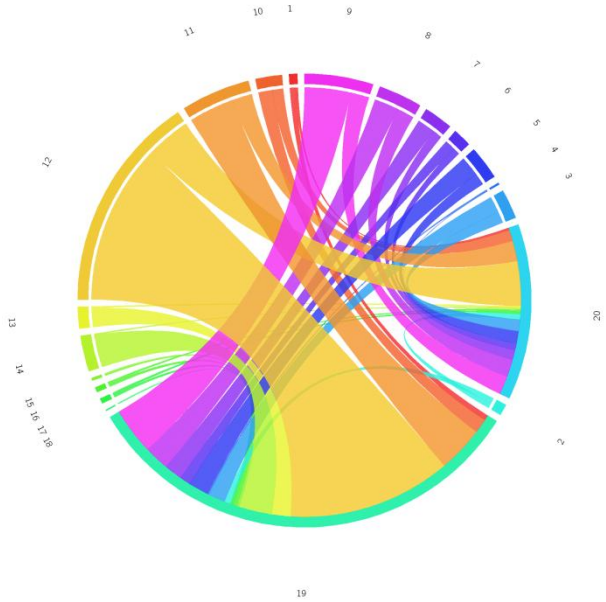


Figure 1: Circos graph - colors represent categories in a table.

	1	2	3	…	10	11	12	…
19	585	605	1740	…	1460	4750	16855	…
20	240	535	1190	…	1170	2090	4620	…

Table 1: Table data used in evaluations (see section 4).

categories 12 and 19, indicated by a large colored ribbon. Ribbons are sized and layered according to data values, such that large values are indicated by large ribbons on top. This paper will investigate an auditory display of Circos - specifically an interactive sonification termed "CircoSonic."

Before discussing CircoSonic, a background of the literature will be presented in section 2, to identify established and unexplored areas of research. Section 3 will outline the development of CircoSonic from the parsing of information to its interactive sonification. Section 4 will discuss the application of the CircoSonic system to a dataset. Sections 5 will describe the evaluation methodology used to compare the performance of Circos, CircoSonic, and a muted version of CircoSonic. Lastly, sections 6 and 7 will present the results and conclusions of the evaluation.

2. BACKGROUND

This section will review the literature to identify existing, emerging and unexplored areas of research. The areas include visualization and genomics, sonification and genomics, sonification of graphs and spreadsheets, and sonification and gaming.

2.1. Visualization and genomics

The field of comparative genomics commonly deals with datasets of chromosome and genome sequences, which are large but static datasets. These can be analyzed by comparing pairings of data values through visualization. An exemplary example is Krzywinski's Circos [12] which progresses from the conventional straight bar diagrams to a circular ideogram and has become a visualization tool useful to other data domains.

"GenomePixelizer" [17] is a visualization tool that allows comparison between more than one pair of genomes. It stacks horizontal bars in a 2D viewer and links duplicate genomic regions with colored lines. ChromoWheel [18] operates as an internet browser application and enables comparisons of multiple genomes similar to GenomePixelizer. Unlike GenomePixelizer it uses a circular layout and draws links that span the interior of the circle. This prevents connecting lines from intersecting other lines and labels.

Circos, appearing as early as 2007, goes beyond ChromoWheel by drawing ribbons instead of lines. This is a small but significant change because it offers an additional dimension to relationships between categories. Circos is a visualization tool that facilitates "the identification and analysis of similarities and difference" [12]. Whilst ChromoWheel and GenomePixelizer simply identify relationships, Circos identifies and provides a visual sense of magnitude for each relationship.

The utilization and development of Circos is continued by "Circoletto" [19] which is "an online visualization tool based on Circos" that offers functionality of the "Basic Local Alignment Search Tool (BLAST)" [20] and supports calculation of sequence similarities, before presenting them visually in a Circos graph.

"Gremlin" [21] goes further to identify an issue and present a solution for Circos's inability to accurately enable "spatial comparisons across rings of varying radii." It demonstrates that the conventional straight bar diagram is more effective for this task, despite connection lines intersecting each other and producing "visual artifacts." Ekdahl [18] in fact recognized the advantages of both the straight bar diagram and circular ideogram. "The circular layout of chromosomes is advantageous for showing relationships between different chromosomes, as the connecting line never crosses over…While the straight bar representation is popular for showing distributions or populations of objects on a chromosome." [22]

Despite recognizing Circos's short comings, O'Brien [21] still states that Circos is "state-of-the-art in genomic rearrangement visualization." Other visualization tools in the comparative genomics field include NCBI map viewer, TIGR Genome Browser, MIPS Arabidopsis Redundancy Viewer and "gff2ps" [17] and Worm-Base (AceDB) cited by [17, 18].

As the visualization strategies developed by the field of comparative genomics are reapplied to other disciplines, there is a need to consider models that deal with dynamic datasets. The examples in this area lack a dynamic dimension because they have not yet needed to deal with them. There is also little research focused on interactivity beyond simply inputting data and navigating the visualizations. There is scope for research looking into a higher degree of interaction including user rearrangement and remapping of the visualization to draw comparisons between component data. Lastly there is opportunity to represent these datasets and their visualizations using sound. There has been some research done in this area, and will be discussed in section 2.2.

2.2. Sonification and genomics

It has been stated in the Sonification Report [3] that projects such as the Human Genome Project require ways to manage and explore the large datasets they collect. Within the field of auditory displays, there has been some research in parametric mapping sonification (PMS or PMSon) as a means to explore data of genomes, proteins and DNA sequences.

Won's [23] sonification experiment of human chromosome 21, sonifies the presence of CpG islands, "because they indicate areas of interest along the genome." This technique is quite specific and cannot be reapplied to other fields without significant modification. Dunn and Clark [24] similarly experimented with a sonification process specific to DNA sequences, proteins and amino acids. Their application of Morse code is very specific for representation of the English alphabet and again cannot be reapplied other data types.

These approaches to genomic sonification have, like the genomic visualization, been domain specific without a generic reapplication to other areas. Circos, although developed by and for the comparative genomics field, has been reapplied successfully to other areas but remains a solely visual form of representation. There has not been any research, to this author's knowledge, that investigates the sonification of Circos.

2.3. Sonification of graphs, charts, spreadsheets and tables

Since "Circos is general and useable in any data domain" [12] a sonification of Circos should consider sonification research that is general and useable in any data domain. The sonification of graphs, charts, spreadsheets and tables are important because unlike genomic sonification they can be applied and used in any data domain.

In the context of auditory displays and even tactile displays, the most common graph investigated is the line graph. Line graphs and bar graphs are the most common graphs for data visualization and as such their auditory display has been covered extensively [25-31]. A comprehensive summary of design guidelines have been presented by Brown [32].

A less common graph, but of more interest to this paper is the pie graph, since the circular geometry is comparable to Circos. Doush *et al.* [29] present a haptic display of the pie graph; while Franklin and Roberts [33] present a purely auditory approach. Surprisingly the latter demonstrated that a non-spatial display inspired by Morse code achieves better accuracy when compared to a spatial audio display. Doush is

one of the few who investigate comparison of pair wise categories. His force feedback design enables "pair wise comparisons of sections of the [pie] chart…the user can select two sections…and reorder [them] to make the two selected sections adjacent." This interactive rearrangement actually affords comparative sonification of pairs; however it is limited to only one pair at a time. Unlike Doush, Circos visualizes the relationship of all existing pairs. The interactivity employed by Doush is limited to components of the graph, rather than the graph as a whole. Neither of the research by Doush *et al.* or Franklin and Roberts investigates the comparison of multiple pie charts.

Since the data used to generate these simple graphs are usually stored in spreadsheets and tables, it is also important to look at the sonification of spreadsheets and tables. Guidelines for auditory display of tabular data are again presented by Brown [32]. Ramloll *et al.* [34] used musical notes in addition to speech to increase accessibility to numeric information in a table. Stockman [35] discusses the lack of accessible spreadsheet applications and existing screen readers that are commonly used to increase accessibility. Stockman's work effectively compliments speech readers by sonifying numeric values. Stockman [36] discusses Mansur who sonified 2d line graphs, by mapping the x-axis and y-axis to time and pitch respectively, similar to Walker's Sonification Sandbox [30, 31]. Stockman [36] concludes that the "interactive control of the sonification can be considerably improved by removing the reliance on CSOUND and generating all sonifications using pre-recorded sounds." Electing to not synthesize sound and use pre-recorded sound for interactive purposes would enable real-time interaction without latency. This is currently how many game engines render audio, primarily to maintain real-time interaction.

2.4. Sonification and gaming

The potential for computer games to contribute to the field of sonification has already been argued by Coleman [37] who found that sound design is highly collaborative and instrumental to the computer game development process. This is specific to Computer game development rather than modification. The latter is an accessible, low budget solution that requires fewer resources such as time, training and finance. The disadvantage, however is that major customization of the game engine itself is not possible without expertise. In contrast to modifying existing game engines, a ground up approach aims to build a tool customized for sonification. Barri's Versum [38] is an example of a ground-up development where a 3D interactive, visual and aural environment was created for sound sequencing. Versum uses Java, SuperCollider and Max/MSP to achieve what closely resembles a gaming engine without a design orientated editor.

Grimshaw [39] conceptually compares a First Person Shooter gaming engine to a sonification system conveying player interaction. Furthermore the potential modification of existing computer game engines for the purpose of sonification has been explored [40, 41]. Game engines have been recognized for their potential to offer real-time collaborative virtual environments [42] using both visualization and "auralization." Both Grimshaw and Le Groux [40] use the Torque Engine while Nguyen [41] uses CryEngine.

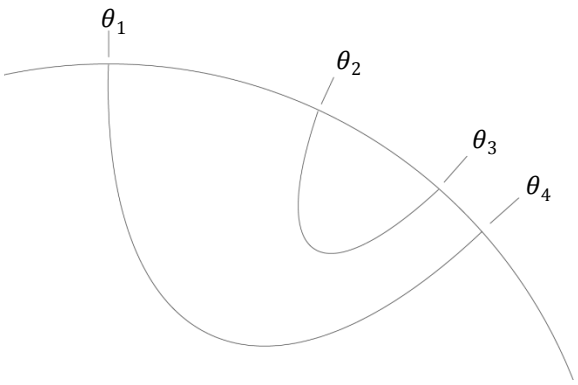


Figure 2: Four angles needed to draw a ribbon.

Many game engine developers offer a level editor, sandbox or toolkit that enables interaction in 3D virtual environments, interaction with real-time dynamic data streams and multimodal feedback. The next section will discuss in detail the use of a gaming engine as a sonification tool.

3. DEVELOPMENT

This section will outline in detail the development of the CircoSonic system. Areas covered in this section include data preparation, drawing the graph, sound parameters, interactivity and sonification.

3.1. Data preparation

The software used is Crysis Wars Sandbox 2, which implements Crytek's CryEngine 2 [http://crytek.com]. Coupled with the FGPS [http://fgps.sourceforge.net] the game engine is capable of reading XML format. Tabular data from a spreadsheet application, such as Excel needs to first be converted to XML. The XML file is read by Crysis Wars Sandbox 2 and each cell value is stored as a variable in game. When two categories in the table are paired (e.g. column/row 4, row/column 2), four angles are calculated to draw the labels and ribbon (see Fig. 2).

3.2. Drawing Circos in Crysis

A Circos graph consists of geometric components such as labels and ribbons; design components such as spacing, color and transparency; and text components such as category headings (see Fig. 3). Using a game engine allows a Circos graph to be drawn in real time from an external XML file. Although this paper discusses CircoSonic's application to static datasets for the purpose of comparative evaluation, its application to dynamic datasets is planned in future work.

The labels around the perimeter are constructed by spawning thick arcs, which include a tick mark with specified translation and rotation in 3D space. Labels draw to the nearest degree and labels smaller than a degree are not drawn. Label headings are drawn as text objects adjacent to each label.

Ribbons link two labels and identify a relationship. They are constructed by stacking thin arcs that are drawn by

spawning a template arc (90 degree arc). The template arc is positioned and rotated before being scaled it in the local x-axis (span) and local y-axis (height).

Colors of both labels and ribbons are selected from a prepared color palette. Colors of labels can either be assigned chromatically or diversely. Ribbon colors are assigned according to the dominant label's color, which is found by comparing the size of a ribbon's two labels. Spacing is added between each row-set for readability and can be specified in numerical units or degrees of rotation. Transparency is uniformly adjustable for all colors and allows readability of intersecting ribbons.

Drawing these in a virtual 3D environment allows multiple Circos graphs to be drawn and overlaid. Circos isolated only allows a side by side comparison of graphs. In the next section the interaction and sonification of a stacked set of Circos graphs will be discussed.

3.3. Interaction and sonification

CircoSonic's sonification is dependent on user interaction. One Interaction, namely rotation, directly affects the sonification by exciting sound. Whilst other interactions such as toggling spin speed, selecting octave and mapping method, indirectly affect the sonification by defining the parameters for selecting what sound to play.

When a user rotates a graph, each label is sonified as it touches a virtual needle fixed at twelve o'clock. The size of the label determines the value to be sonified whilst the user defined parameters determine how the value is sonified.

3.4. Sound parameters and preparation

The static sound parameters include timbre and volume, whilst the dynamic sound parameter is limited to tone. The user defined parameters include octave, tempo and mapping method. All sounds are musical tones of the western chromatic scale and were generated from MIDI before being compressed as an FMod library [http://www.fmod.org] for compatibility with the game engine. This strategy of pre-recording sounds affords real-time interactivity without latency.

3.5. Keyboard interaction

Users can rotate each circular graph using the num-pad keys on a keyboard. The three graphs can be rotated separately or collectively. The speed of rotation is toggled using the "shift" key, holding down to increase speed and releasing to decrease speed. With increased spin speed, the tempo of the sonification provides an overall sense of the dataset. With decreased spin speed, the detailed sections of the data can be interrogated more closely. The zoom is changed by using the "plus" and "minus" keys, which moves the camera position respectively closer or farther from the graphs.

3.6. Mouse interaction and mapping methods

A mouse enabled text based interface (see Fig. 3) allows the user to define parameters which affect the sonification. The "active" check box allows users to select which graphs are

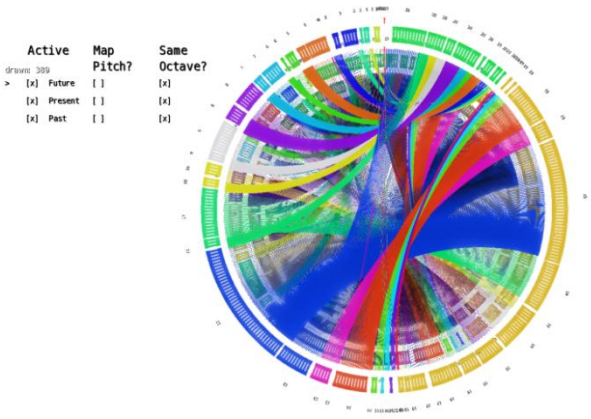


Figure 3: CircoSonic in a modified computer game engine.

revealed or hidden. This gives users an option to reduce visual loading to aid cross graph comparison.

The user can select from two mapping methods to determine the musical tone to play. The first method maps values to a linear but inverse progression of chromatic tones. A high value will sound a low tone, and a low value will sound a high tone. The second method maps values to tonality in accordance with [41, 43] in which the circle of fifths is used to determine a non-linear spectrum of tones. The aim of this method is to allow users to listen to what sounds "out of key". For example a C4 and a C4 will indicate values of no variance, whilst a C4 and F#4 will indicate values of maximum variance; furthermore a C4 and a G4 will indicate values of minimal variance.

User can also select, using another check box, to play graphs in the same octave. When selected, all graphs will play in a middle octave. When deselected, each graph will play in a different octave; the top graph in a high octave and the bottom graph in a low octave.

These user defined parameters ultimately affect how the sonification will sound and can be changed real-time whilst rotating a graph.

3.7. Timbre and volume

The timbre used was concert grand piano and the volume was set to a fixed dB. Volume did change dynamically as a consequence of user interaction. As a Circos graph is rotated quickly, any similar data values will play the same tone. The effect is multiple sound sources playing the same tone which seemingly increases volume. This enables users to use volume as an indicator of data point frequency.

CircoSonic has the ability to represent data in both visual and aural forms which identifies relationships and convey their biasness. CircoSonic only sounds upon excitation through user interaction and can be used to compare multiple Circos graphs. The next section will discuss the application of CircoSonic to a real dataset.

4. APPLICATION

The developed system CircoSonic was applied to a dataset of historic, current and projected water availability of the Murray Darling Basin (MDB). The MDB is the catchment system serving the largest river in Australia, the Murray-Darling River.

The MDB dataset used is publically available [44]. The data is in the form of a table and presents eight cases in various stages of development (without development, current development, future development) and climate (historical, recent, wet, dry and median 2030), which are further broken down into sub-categories (water inflow, losses, end flow, diversions, groundwater losses, average surface water available, and relative level of surface water use – all given in giga-litres per year except for the last which is given as a percentile). For this paper only three of eight cases have been selected: (1) historical climate without development, (2) historical climate with current development, (3) projected climate for 2030 with future development. The water inflows (see Fig. 1, category 19) and losses (Fig. 1, category 20) of the 18 catchments (Fig. 1, categories 1-18) are transferred to a separate spreadsheet in preparation for importation into the game engine.

A demonstration of this application is included in the supplementary materials as videos displaying the sonification and its interactivity. In the next section the evaluation will compare three systems using the same MDB dataset.

5. EVALUATION

The method of evaluation will be outlined in this section. See section 6 for discussion of the outcomes.

An insight based methodology is used to evaluate CircoSonic, similar to [21] in which Circos was compared to Gremlin. By employing this methodology a direct comparison between Circos and CircoSonic is established, and an indirect comparison of CircoSonic to Gremlin is accommodated. A muted version of CS was included to establish a comparison between visualizations, from which a comparison between visualization and sonification could be extrapolated.

5.1. Insight based methodology

An insight based methodology [45, 46] quantifies the performance of a system based on qualitative insights generated by a participant using the system. In this case the three systems being compared are Circos (C), CircoSonic (CS), and CircoSonicMuted (CSM). The Circos graphs evaluated were generated using table viewer (see Fig. 1).

In accordance with [21], an "insight" is defined to be "a unique, individual observation about the data by a participant" and can be further categorized by complexity:

Type A: Simple - discernible from textual analysis.

Type B: Detailed - not readily apparent through textual analysis.

Type C: Detailed Contextualization - involving cross-referencing of observations or knowledge base.

5.2. Hypothesis

The hypotheses for the evaluation comparing C, CS and CSM are:

H.1: CS will outperform C at generating a higher average number of (a) total insights and (b) type C complex insights.

H.2: CS will outperform CSM at generating a higher average number of (a) total insights and (b) type C complex insights.

5.3. Pilot evaluation

The pilot evaluation analyzed insights per second over two 5-minute sessions, however it became apparent that this awarded an undue bias towards the non-interactive visualization since it was less time sensitive than the interactive sonification. Listening to and interacting with CS and CSM required an investment of time which effectively reduced the rate of generated insights, whilst potentially increasing the end total of generated insights during an unrestricted session. For this reason, the sessions were re-conducted in the final evaluation as open-ended sessions (see section 5.5).

5.4. Participants

There were eight participants including a mixture of female and male, Master graduates and PhD students (see Table 2). None had eyesight or hearing impairments and all demonstrated simple comparative pitch and volume recognition. Music expertise was not a requirement since the link between music expertise and performance of sound perception tasks has not yet been established [47]. All eight had little to no experience with both Circos and CircoSonic. Some had previously been exposed to the dataset. Each of the participants was allocated a group number that determined the order in which they would use each system (see Table 3).

	sex	edu	group	data familiarity
P1	f	PhD	1	yes
P2	m	M	1	yes
P3	m	PhD	2	yes
P4	f	M	2	yes
P5	f	M	3	yes
P6	f	M	3	no
P7	m	M	4	no
P8	f	M	4	no

Table 2: Participants (P1-8).

5.5. Session protocol

Each participant performed consecutive sessions in which they were exposed to two of the three systems (C, CS, CSM). Each included (a) 15 minutes tutorial and explanation, (b) an open-ended session using one system, and (c) an open-ended session using a different system. The tutorial covered how to read C and listen to CS, in that respective order, and used example datasets unrelated to the datasets given in the sessions. The

explanation covered the format of the sessions, background on the dataset and instruction to make observations during the sessions by thinking aloud. Each session was recorded on audio with the consent of participants and concluded when the participant stated they could not make any more observations. The order in which participants were exposed to each system considered a potential learning curve and the effects of fatigue (see Table 3).

group	1st	2nd
1	C	CS
2	CS	C
3	C	CSM
4	CSM	C

Table 3: The order in which each group used the systems.

5.6. Assessing insights

Observations made by participants were assessed against the definition of an "insight" (see section 5.1). Insights were categorized by complexity into type A, B and C and quantified by counting.

The typical type A insights included the identification of size differences or similarities over the 3 graphs, the recognition of ordering, and the recognition of biasness or equality between values. Typical type B insights included the recognition of changes to ordering, recognition of changes to biasness and articulating the ratio of biasness. Typical type C insights were limited to conclusions drawn by cross-referring the above types or using their knowledge base to contextualize the information.

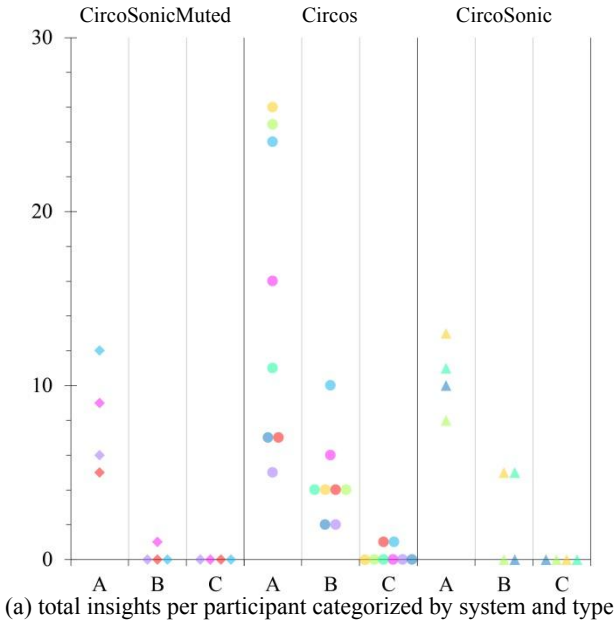
6. RESULTS

The results of the evaluation are presented in Fig. 4. All charts show CSM, C and CS respectively from left to right. Fig. 4 (a) shows the number of insights made by participants, separated into type A, B and C insights. Fig. 4 (b) shows the total insights of each participant with the averages indicated by a cross. Each participant is represented with a different color that corresponds between Fig. 4 (a) and (b). Lastly, Fig. 4 (c) shows the average component breakdown of categorized insights.

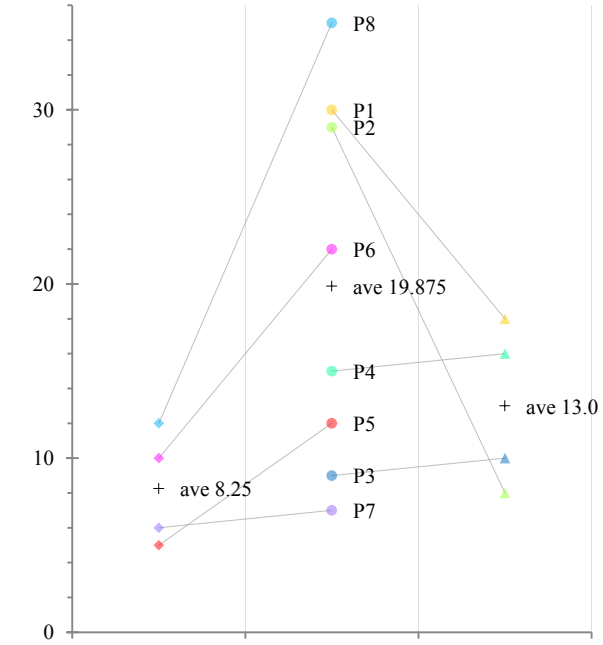
The results show that across all three systems C outperformed both CS and CSM, despite C being non-interactive and solely visual. C achieved a higher number of total insights and a higher number of insights per category for type A, B and C.

There were two participants who performed better on CS than C, which is a marginal but promising result. These two participants were members of group 2, which may suggest that participants who used CS first were subject to more fatigue than their counterparts in group 1.

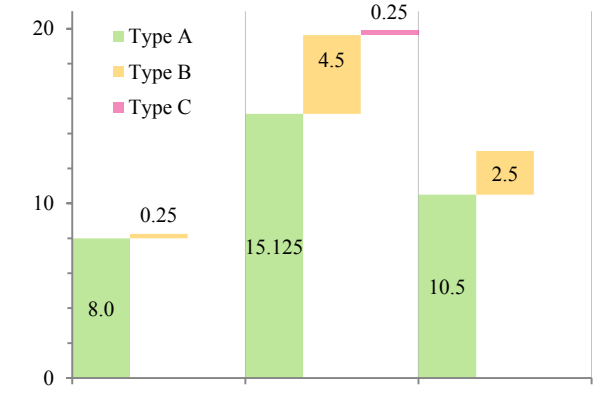
For all participants, type A insights were the most common and type C were the least common. This is in line with expectations since all participants were equally inexperienced at C, CS and CSM. Even though both CS and CSM failed to generate any type C insights, CS did consistently outperform CSM at generating a higher number of type A and B insights and consequently a higher number of total insights.



(a) total insights per participant categorized by system and type



(b) total insights per participant categorized by system



(c) average insights per system categorized by type

Figure 4: Results of evaluation.

6.1. Discussion of insights

The most common insight made was identifying maximum value. This was actually maximum value proportionally to the whole set, since the size of a ribbon is relative to the whole circle. Minimum value was also recognized, however it was almost always after maximum. This could be an effect of the tutorial which demonstrated how to recognize maximum values before how to recognize minimum values. Both the identification of maximum and minimum values was categorized as a type A insight.

A common type B insight was the recognition of changes in ordering. This was only possible using system C, which automatically reordered the layering of ribbons based on size. CS and CSM did not feature this ordering mechanism. Another common type B insight made between the past, present and future datasets was the recognition that there were significant changes between past and present, and only minor changes between present and future.

There were only two type C insights made. One was the recognition that proposed changes for the future were insufficient to restore historic patterns. The other was the conclusion of a distributed increase to supply the significant loss of inflow into two catchments. These were made by two participants of different groups using C.

7. CONCLUSION

The only hypothesis found to be true was H.2 (a): CircoSonic outperformed CircoSonicMuted at generating a higher number of total insights. Neither CircoSonic nor CircoSonicMuted generated Type C insights (that is context referenced insights, see section 5.1) in the evaluation.

The comparison of Circos to CircoSonicMuted suggests that the visual component of CircoSonic heavily underperformed, limiting its overall performance. The sound component of CircoSonic consistently improved the generation of insights beyond CircoSonicMuted, which demonstrates that CircoSonic with sound performs better than CircoSonic without sound. One of the most significant strengths of Circos is the automated ordering of ribbons based on their size. The difference between the orders in which ribbons are layered is clearly noticeable and generates more insights as a consequence. CircoSonic's layering and transparency needs to be developed further if its visualization is to perform as well as Circos.

The results of the evaluation suggest that when representing a static dataset for the purpose of data mining, a static visualization is more appropriate than an interactive visualization/sonification. An evaluation involving dynamic data may show different results, however Circos does not currently support dynamic datasets.

8. FUTURE WORK

This paper has presented the development of an interactive system to explore table datasets through visualization and sonification. It has been applied to a static dataset, however it is planned that the same system be applied to a dynamic dataset. There is currently a gap between Excel and Crysis which requires data to be reformatted into a custom XML format. It is planned that this gap will be filled by reading directly from Excel via Open XML format. There is also scope to explore sound parameters such as timbre, and mapping methods such as frequency and volume. The positioning of CircoSonic within a gaming engine also lends itself to be extended to an ambisonic or collaborative/interactive system.

CircoSonic is currently being applied to pedestrian movement and natural surveillance in the field of architecture. The keyboard and mouse interactivity presented in this paper has since been developed further to include the ability to control rotation using the Apple iPhone and UDP.

The novel visualization Circos is but one emerging scientific visualization that requires investigation from the perspective of auditory displays. The investigation of novel visualizations from the perspective of auditory displays needs to be extended to include those which deal with multivariate and dynamic datasets whilst still offering a broader application to diverse data domains.

9. ACKNOWLEDGMENTS

This research is jointly funded by the Australian Research Council (ARC LP 0991589), the University of New South Wales (UNSW), and the Emergency Information Coordination Unit (EICU). The author wishes to thank Tim Stubbs *et al.*, for informative discussion regarding the Murray Darling Basin at the Data Visualization Workshop, UNSW, Sydney, Nov 2011.

10. REFERENCES

[1] E. Tufte, *Envisioning Information*, Cheshire, Connecticut, USA: Graphics Press, 1990.

[2] E. Tufte, *The Visual Display of Quantitative Information*, Cheshire, Connecticut, USA: Graphics Press, 1983.

[3] G. Kramer, B. Walker, T. Bonebright *et al.*, *The Sonification Report: Status of the Field and Research Agenda*, Prepared for the National Science Foundation by members of the Int. Community for Auditory Display, 1999.

[4] D. Smith, and B. Walker, "Tick-marks, axes, and labels: The effects of adding context to auditory graphs," in *Proc. of the 8th Int. Conf. on Auditory Display*, Kyoto, Japan, 2002.

[5] K. Nesbitt, "Comparing and reusing visualisation and sonification designs using the MS-taxonomy," in *Proc. of the 10th Int. Conf. on Auditory Display*, Sydney, Australia, 2004.

[6] B. Walker, and G. Kramer, "Mappings and metaphors in auditory displays: an experimental assessment," *ACM Trans. on Applied Perception*, vol. 2, no. 4, pp. 407-412, October, 2005.

[7] C. McCormick, and J. Flowers, "Perceiving the relationship between discrete and continuous data: a comparison of sonified data display formats," in *Proc. of the 13th Int. Conf. on Auditory Display*, Montréal, Canada, 2007.

[8] K. Beilharz, and S. Ferguson, "An interface and framework design for interactive aesthetic sonification," in *Proc. of the 15th Int. Conf. on Auditory Display,* Copenhagen, Denmark, 2009.

[9] B. Walker, and M. Nees, "Theory of Sonification," *The Sonification Handbook*, T. Hermann, ed., pp. 9-40, Berlin, Germany: Logos Verlag, 2011.
[10] S. Ferguson, W. Martens, and D. Cabrera, "Statistical Sonification for Exploratory Data Analysis," *The Sonification Handbook,* T. Hermann, ed., pp. 175-196, Berlin, Germany: Logos Verlag, 2011.
[11] E. Brazil, and M. Fernström, "Auditory Icons," *The Sonification Handbook*, T. Hermann, ed., pp. 325-338, Berlin, Germany: Logos Verlag, 2011.
[12] M. Krzywinski, J. Schein, I. Birol *et al.*, "Circos: an information aesthetic for comparative genomics," *Genome Research*, vol. 19, no. 9, pp. 1639-1645, 2009.
[13] D. Constantine, "Close-ups of the genome, species by species by species," *New York Times*, pp. F4, 23 Jan, 2007.
[14] D. Duncan, "Welcome to the future," *Conde Nast Portfolio*, pp. 192-197, 220-222, November, 2007.
[15] E. Ostrander, "Genetics and the shape of dogs," *American Science*, vol. 95, pp. 406-413, 2007.
[16] C. Zimmer, "Now: the rest of the genome," *New York Times*, pp. D1, 11 November, 2008.
[17] A. Kozik, M. A. Marra, and R. Michelmore, "GenomePixelizer - a visualization program for comparative genomics within and between species," *Bioinformatics*, vol. 18, no. 2, pp. 335-336, 2002.
[18] S. Ekdahl, and E. Sonnhammer, "ChromoWheel: a new spin on eukaryotic chromosome visualization," *Bioinformatics,* vol. 20, no. 4, pp. 576-577, 2004.
[19] N. Darzentas, "Circoletto: visualizing sequence similarity with Circos," *Bioinformatics*, vol. 26, no. 20, pp. 2620-2621, 2010.
[20] S. Altschul *et al.*, "Basic local alignment search tool," *J. Molecular Biology*, vol. 215, pp. 403-410, 1990.
[21] T. O'Brien, A. Ritz, B. Raphael *et al.*, "Gremlin: an interactive visualization model for analyzing genomic rearrangements," *IEEE Trans. on Visualization and Computer Graphics*, vol. 16, no. 6, pp. 918-926, 2010.
[22] T. Harris *et al.*, "WormBase: a cross-species database for comparative genomics," *Nucleic Acids Research,* vol. 31, pp. 133-137, 2003.
[23] S. Won, "Auditory display of genome data: human chromosome 21," in *Proc. of the 11th Int. Conf. on Auditory Display*, Limerick, Ireland, 2005, pp. 280-282.
[24] J. Dunn, and M. Clark, ""Life Music": the sonification of proteins," *Leonardo,* vol. 32, no. 1, pp. 25-32, 1999.
[25] R. Ramloll, W. Yu, S. Brewster *et al.*, "Constructing sonified haptic line graphs for the blind student: first steps," in *Proc. of the 4th Int. ACM Conf. on Assistive Technologies,* Arlington, USA, 2000.
[26] J. Roberts *et al.*, "Virtual haptic exploratory visualization of line graphs and charts," in *Proc. of the Conf. onStereoscopic Displays and Virtual Reality Systems IX,* San Jose, CA, USA, 2002, pp. 401-410.
[27] L. Brown, S. Brewster, R. Ramloll *et al.*, "Browsing modes for exploring sonified line graphs," in *Proc. of the 16th British HCI Conf.*, London, UK, 2002, pp. 6-9.
[28] L. Brown, and S. Brewster, "Drawing by ear: interpreting sonified line graphs," in *Proc. of the 9th Int. Conf. on Auditory Display,* Boston, USA, 2003.
[29] I. Doush *et al.*, "Making Microsoft Excel: multimodal presentation of charts," in *Proc. of the 11th Int. Conf. on Computers and Accessibility (ACM SIGACCESS),* Pittsburgh, USA, 2009.
[30] B. Walker, and J. Cothran, "Sonification Sandbox: a graphical toolkit for auditory graphs," in *Proc. of the 9th Int. Conf. on Auditory Display*, Boston, USA, 2003.
[31] B. Davison, and B. Walker, "Sonification Sandbox reconstruction: software standard for auditory graphs," in *Proc. of the 13th Int. Conf. on Auditory Display,* Montreal, Canada, 2007, pp. 509-512.
[32] L. Brown *et al.*, "Design guidelines for audio presentation of graphs and tables," in *Proc. of the 9th Int. Conf. on Auditory Display,* Boston, USA, 2003, pp. 284-287.
[33] K. Franklin, and J. Roberts, "Pie chart sonification," in *Proc. of the 7th Int. Conf. on Information Visualization,* 2003, pp. 4-9.
[34] R. Ramloll, S. Brewster, W. Yu *et al.*, "Using nonspeech sounds to improve access to 2D tabular numerical information for visually impaired users," in *Proc. of the BCS IHM-HCI,* Lille, France, 2001, pp. 515-530.
[35] T. Stockman, "The design and evaluation of auditory access to spreadsheets," in *Proc. of the 10th Int. Conf. on Auditory Display*, Sydney, Australia, 2004.
[36] T. Stockman, "Interactive sonification of spreadsheets," in *Proc. of the 11th Int. Conf. on Auditory Display,* Limerick, Ireland, 2005.
[37] G. Coleman *et al.*, "Approaches to auditory interface design - lessons from computer games," in *Proc. of the 11th Int. Conf. on Auditory Display*, Limerick, Ireland, 2005, pp. 99-104.
[38] T. Barri, "Versum: audiovisual composing in 3D," in *Proc. of the 15th Int. Conf. on Auditory Display,* Copenhagen, Denmark, 2009.
[39] M. Grimshaw, "Sound and immersion in the first-person shooter," in *Proc. of the Int. Conf. on Computer Games (CGames)*, La Rochelle, France, 2007.
[40] S. LeGroux *et al.*, "VR-RoBoser: real-time adaptive sonification of virtual environments based on avatar behavior," in *Proc. of the Conf. on New Interfaces for Musical Expression*, New York, USA, 2007, pp. 371-374.
[41] V. Nguyen, "Tonal DisCo: dissonance and consonance in a gaming engine," in *Proc. of the 17th Int. Conf. on Auditory Display*, Budapest, Hungary, 2011.
[42] J. Moloney, and L. Harvey, "Visualization and 'auralization' of architectural design in a game engine based collaborative virtual environment," in *Proc. of the 8th Int. Conf. on Information Visualisation*, 2004, pp. 827-832.
[43] S. Malinowski. "Harmonic Coloring Based on the Perfect Fifth," http://www.musanim.com/mam/pfifth.htm.
[44] CSIRO, "Water Availability in the Murray-Darling Basin," 2008. See "Appendix A: Surface water availability and use," pp. 59, http://www.csiro.au/files/files/po0n.pdf.
[45] P. Saraiya, C. North, and K. Duca, "An insight-based methodology for evaluating bioinformatics visualizations," *IEEE Trans. on Visualization and Computer Graphics,* vol. 11, no. 4, pp. 443-456, 2005.
[46] C. North, "Toward measuring visualization insight," *IEEE Computer Graphics and Appl.*, vol. 26, no. 3, 2006.
[47] T. L. Bonebright, and J. H. Flowers, "Evaluation of Auditory Display," *The Sonification Handbook*, T. Hermann, ed., pp. 111-144, Berlin, Germany: Logos Verlag, 2011.

TWEETSCAPES – REAL-TIME SONIFICATION OF TWITTER DATA STREAMS FOR RADIO BROADCASTING

Thomas Hermann[1], *Anselm V. Nehls*[2], *Florian Eitel*[3], *Tarik Barri, Marcus Gammel*[4]

[1] Ambient Intelligence Group, CITEC, Bielefeld University, Bielefeld, Germany
[2] HEAVYLISTENING , Berlin, Germany
[3] Freelance programmer, Berlin, Germany
[4] Deutschlandradio Kultur, Berlin, Germany
thermann@techfak.uni-bielefeld.de

ABSTRACT

This paper introduces *tweetscapes*[1], a system that transforms message streams from Twitter in real-time into a soundscape that allows the listener to perceive characteristics of Twitter messages such as their density, origin, impact, or how topics change over time. *Tweetscapes* allows the listener to be in touch with the social platform/medium *Twitter* and to understand its dynamics. We developed *tweetscapes* with and for the Sound Art department of the Germany-wide radio program *Deutschlandradio Kultur* where the sonifications are now broadcasted several times per week for a few minutes since October 2011. The goal was to create a new sense of media awareness and an example of how sound can support monitoring applications differently than mere alarms. This paper introduces the methods, the ideas, the design, the sounds, and it discusses our experiences with, and novel interaction possibilities offered by *tweetscapes*.

Figure 1: Screenshot of the *#tweetscapes* media stream at http://tweetscapes.de (2012-01-18): arrows show replies, #hashtags occur at the location of the tweet.

1. INTRODUCTION

One of the major advantages of sonification is that it enables the communication of information without requiring any visual attention and thus without any interference with a visual task. This makes sonification not only highly attractive for process monitoring tasks, (see [1]), where a process is to be followed while engaged into another primary task, or for information displays for the visually impaired who cannot access any visual information (e.g. see [2]), but also for radio broadcasts where there is simply no visual channel.

Since sonification can convey complex and detailed information, and we live in a decade of steadily growing information spaces, it is astonishing that it is nowadays so rarely used in established radio formats. To our knowledge the first regular use of sonification in a radio program was *broadcasting auditory weather forecasts*[2], a system introduced in [3] that represented many details of the expected weather (e. g. temperature, humidity, precipitation, wind, etc.) and its expected change over time for the next 24 hours in a 12 seconds soundscape, tuned to convey quickly and without the detour via language processing a good impression of how the weather is going to be like. From that project we learned that sonification in radio faces the particular challenge that sound needs to be as self-explanatory as possible and that the sonifications will be heard in many different contexts such as in the car, during work, in noisy environments – which imposes specific constraints on the sonification design.

As partnership and cooperation between Deutschlandradio Kultur and the Ambient Intelligence group at CITEC, we decided to create a new series called *Sonarisations*, where sonifications will be featured within the nationwide radio program *Deutschlandradio Kultur*. The given way of embedding the sonifications into the program – as gap filler between broadcasts and the news – provided some constraints for the selection of the domain as outlined in detail in Section 8. Furthermore we agreed that the tight cooperation of sonification

[1] offical name: '*#tweetscapes*'; we omit the '#' to increase readability
[2] German title: 'Die Wettervorhörsage'

scientists and artists/sound designers would be required.

Tweetscapes is the first and pilot project to establish and kick-off the series of Sonarisations. We conducted a workshop and presentation with support from Sam Auinger and Martin Supper at UdK Berlin (sound studies/acoustic communication) and decided subsequently to follow the second authors' proposal to create a real-time sonification of Twitter traffic. The proposal was then jointly elaborated in tight dialogue between the involved artist and sonification scientist, the process, interesting in itself and discussed in [4], will only be referred to occasionally in this paper. The resulting soundscape aimed to be both aesthetically interesting and useful as a sonification, i.e. key principles for sonification such as reproducibility, precise algorithmic transformation [5] are respected.

Twitter serves as a good example for communication networks where complex interactions between agents have shifted from the real world to the virtual/digital realm; as a whole the network shows an an overall behavior which is difficult to grasp, if at all, from merely looking at few tweets. How do individual messages lead to tweet avalanches which become trending topics? How does the Twitter community respond to events in the worlds, ranging from simple events like the onset of advertisement breaks in the big German TV shows to breaking news? How can sound provide a new level of experience of the digital medium? and how can we best make sonification more widely known and accepted as medium? *Tweetscapes* follows these questions and furthermore showcases an interdisciplinary experiment between media, auditory display research and sound arts. This paper aims at explaining the sonification side as the main focus, but the other aspects will be touched on as well.

We start with a short introduction into the social communication medium Twitter and summarize the key phenomena that we find relevant. This leads us in Section 3 to the goals and design ideas of *tweetscapes*. In Section 5 we introduce the sonification methods stream by stream. For the website, we worked on an audiovisual stream (Section 6) where the synchronization of visual and auditory events helps to better understand the data. Section 7 provides and comments on different *tweetscapes* for typical activity patterns. Finally, we address some practical issues and share our experience when integrating *tweetscapes* to the radio program of Deutschlandradio Kultur.

2. TWITTER — MICRO-BLOGGING DATA STREAMS

Twitter is a social networking service that allows users to send *tweets*: short text messages of up to 140 characters. It has grown since 2006 to a globally known service. Registered users can follow the tweets of other users and thus become 'followers'. Topics are set by using *hashtags* which are simply words prefixed with the # symbol. Instead of watching the posts of users they follow, users can also query the Twitter stream for specific keywords and thus use Twitter as a news filter. According to wikipedia, Twitter has 140 million users[3]. The amount of information per day is incredible and difficult to understand as a whole from the microscopic views that the standard interfaces offer.

3. TWEETSCAPES: GOALS AND DESIGN IDEA

Tweetscapes follows several goals on different levels: from the perspective of sonification research, the goal was to make the idea of sonification more publicly known by integrating it into the regular radio program. From the perspective of radio makers, it should be aesthetically interesting and surprising, and touch a subject that is of public and cultural interest. The real-time sonification of Twitter traffic was a topic that is compatible with these different goals.

The key design idea is to create a soundscape that involves several sound streams, similar to the sound- (or land-) scapes that surround us in real environments. They typically have a foreground, middle- and background. Likewise,

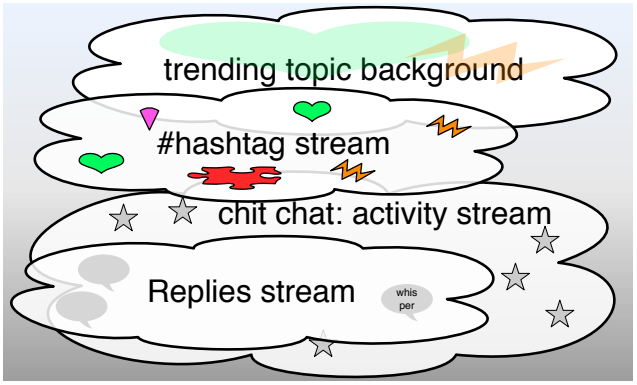


Figure 2: Sonic streams of *#tweetscapes*: salient hashtag events dominate a multi-stream background with activity, replies and topic streams.

tweetscapes represents the Twitter activity in several auditory streams: (a) *chit chat* is a stream where all tweets occur that are neither replies nor have hashtags, (b) *replies* is a stream of sonic events for public tweets exchanged by users, (c) *hashtag events* form the acoustic foreground stream where distinct topics become clearly audible, and finally (d) a *topic stream* makes the three most trending topics continuously perceivable as a background stream.

Apart from (d), all streams consist of individual sound events which are caused by tweets and thus are a true 1:1

[3] on March 21, 2012, see Section on growth on http://en.wikipedia.org/wiki/Twitter

representation of the event-like communication in Twitter. (a), (b), and (c) provide different filters or views. The sound events are chosen from a huge library of with different sonic material of large variation (as explained below) so that the overall sonic shape becomes acoustically rich. Finally the overall activity is estimated by a some features such as the average frequency of tweets. This parameter is used to influence the sound on many levels, such as sound effects, global parameters and post-processing.

4. PRE-PROCESSING OF TWITTER DATA

Twitter can be accessed through the Streaming-API[4]. This returns all tweets matching the particular query parameters in real-time. It is possible to filter by user names, keywords or location. The query is transmitted by a parameter in the HTTP[5] request. Twitter doesn't terminate the connection but sends new, matching tweets in real-time.

Unfortunately, there is no way to query tweets according a particular language. A series of tests showed that only 0.33 tweets per second are labeled with location information. This issue is solved by logging tweets with a very generic search query over a long time period. Based on this data the word frequency is analyzed regarding words from German users and non-German users. This results in a word list filled with words which are mostly used in German language.

Due to performance issues Twitter limits the rate of results on highly general search queries. To cope with these limits and collect nonetheless as complete as possible the German Twitter traffic, we created a restricted word list. The challenge is to filter these and suppress as good as possible the non-German tweets that may appear since words on the list are identical with words in other languages. This is also taken into account with the word list selection.

Every transmitted tweet is encoded as JSON[6] and contains approximately 54 parameters[7], which are related to the tweet or its sender. These characteristics are filtered, processed and enhanced as follows:

The location is important, especially for the visualization (see Section 6, below). If no location is set in the tweet the program takes a guess of coordinates based on location settings in the user preferences. If this is not successful a random position on the German map is created and cached based on user ID for a short time period so that repeated tweets from that user appear at the same location.

The hashtags need particular attention: A counter is incremented for every occurring hashtag h. The relative occurrence estimates the current popularity of the hashtag. The value is updated every 10 s by $m_h = \lambda m_h + (1 - \lambda) N_h$, where N_h is the number of occurrences over the past 10 seconds. We set λ of this 'leaky integrator' to get a half-life value of 5 minutes. This results in a dynamic ranking of all incoming hashtags. A ranking of the top 20 popular keywords is continuously extracted and sent to the visualization and sonification modules.

Additionally many more characteristics are processed, starting from simple metrics such as the number of followers of a user (followers count), retweet count of a tweet or character count of the tweeted message towards more complex parameters such as the time difference between a tweet and

feature	description/ type
realtimestamp	absolute time of tweet (by *tweetscapes*) float, ms since 2011-08-01
created_at	absolute time of tweet (by Twitter) float, ms since 2011-08-01
is_a_reply	flag if tweet is a reply to another integer (0/1)
RT_count	number of retweets integer, upper limit 100
text	chars of tweet text integer
user_followers_count	followers of User integer
user_statuses_count	count of tweets from User integer
RT_created_at	seconds since retweeted status integer, sec (default 0)
RT_statuses_count	count of tweets from retweeted user integer (default 0)
RT_followers_count	count of followers from retweeted user integer (default 0)
weekday	current weekday integer (mon=0)
sec_since_midnight	seconds since midnight integer, sec
mood	mood of tweet (guessed by emoticons) integer
question	number of question marks integer
longitude	longitude of Tweet float (default random)
latitude	latitude of Tweet float (default random)
tophashtag	best rated hashtag used in Tweet string (default " ")
relative_rating	best rated hashtag / current top hashtag float (default 0)
tweet_id	ID of this tweet string
RT_tweet_id	ID of retweeted tweet string (default " ")

Table 1: Extracted Features that characterize tweets in *#tweetscapes*.

[4]https://dev.twitter.com/docs/streaming-api/
[5]http://tools.ietf.org/html/rfc2616
[6]http://www.json.org/
[7]https://dev.twitter.com/docs/api/1/

later retweets, or 'mood detection' via the emoticons contained in tweets. Table 1 gives a complete overview of all extracted features. Finally, and very relevant for *tweetscapes*, statuses and users are filtered by a blacklist and identical tweets are blocked in a given time period to avoid spam.

These preprocessing results in 20 features which are encoded in Open Sound Control[8] (OSC) messages sent to the visualization and sonification modules. To allow multiple applications to access the data (debug, visuals, sound, logging, ...) the stream is not sent – as designed by OSC – over UDP but using a TCP connection. This enables the encapsulation into multiple servers and a clean interface between the different parts. The OSC processing applications usually require UDP packets, so a reliable proxy is used for parsing OSC packets out of the TCP stream and to translate them back to UDP.

5. SONIFICATION METHODS FOR THE *TWEETSCAPES* SOUND STREAMS

We will now discuss the sound streams and explain why and how the tweet features control the parameters of the sound events. Please navigate to http://tweetscapes.de[9] to familiarize yourself via the real-time stream with the sounds. As outlined in Section 3, the sonification contains four sound streams which we introduce next.

5.1. The chit-chat stream

As tweets are events, the most straightforward idea is to take a 1:1 manifestation of tweets as sound events. This resembles the Geiger counter that represents individual radioactive events as sound grains. Likewise a direct event sonification creates perceptual units on a higher level, such as the perception of momentary density and its change, of rhythms and waves. Beyond that, with the event sounds conveying details of the tweets, temporal patterns emerge that may become auditory gestalts. Our first attempt for such a granular texture of event actually used chirped sine tones to create a soundscape similar to literally twittering birds. Two sound examples are provided at our website.[10] Obviously, the bird sounds fill the sound space quite intensively. For that reason we considered other timbre spaces. We finally decided to use highly transient, non-pitched, short sounds. As sound source material we chose 8 sample sets of each 20 sounds from the area of communication, including single typewriter events, computer keystrokes, morse keys and relay clicks. Instead of modulating or manipulating features of single sounds, we decided to start from *ordered set of sounds*, (e.g. keystroke recordings at increasing force) and select the sample to be used according to the tweet's feature value. In this way, we automatically encode a data feature as a coherent auditory unit. For instance, the sample selection is driven by the number of followers of the tweet writer. Since tweets have obviously a higher impact depending on that feature, this 'impact' becomes literally perceivable as keystroke impact, which manifests in correlated level, brightness, complexity, duration etc. Technically this method can be regarded as a *parameterized auditory icon* [6] approach, yet the parameterization is here not achieved by a complex synthesis but via a table look-up. The term 'Sound Font' can be used for this battery of ordered samples.

In a nutshell the mapping[11] to sonic features is:

- impact (couples attack, level, timbre, etc., achieved via data-driven sample selection in ordered sample set [0,19]) ← user_followers_count.
- stereo panning [left, right] ← longitude [eastern, western edge of Germany], i.e. stereo position is as if the listener would be located in the center of Germany.
- reverberation [dry, wet] ← distance [0, 1000 km] from the center of Germany
- delay time decreases, and delay feedback increases with increasing RT_count, so that retweets can be recognized by their echo effect.
- sample set selection [complex, tiny] ← global activity [low, high], i. e. during lower activity the higher sparseness allows the program to select more complex sounds.

The algorithm is prepared to work with N-channel audio systems so that beyond a stereo panning also the latitude is properly mapped. Sound example S3 demonstrates chit-chat events for two single tweets, one near east, the second far away in the south. S4 contains two retweets, the first with RT_count $= 30$, the second with > 100. The spatial drift represents the spatial difference between the original tweet and the retweet location. Sound example S5 contains 5 selected chit-chat events with increasing impact (i. e. user_follower_count) Finally, sound example S6 is a typical chit-chat stream for German Twitter traffic.

5.2. The Replies sound stream

Replies are part of the public conversation at Twitter, but they are usually directed at a specific person. They should stand out of the chit-chat stream and have their own character and timbre so that listeners can perceive the ratio of non-replies tweets to replies from their occurrence frequency. A good metaphor is that of whispering. Similarly to the sound font

[8] http://opensoundcontrol.org/

[9] english version at http://tweetscapes.de/?lang=en

[10] http://techfak.uni-bielefeld.de/ags/ami/publications/HNEBG2012-TRT

[11] reported as sound parameter [min., max.] ← data feature [min., max.], using a linear mapping if not otherwise stated.

approach for chit-chat, here some longer samples of whispering are used where the whispering style gets more and more excited and faster with time. The length of a reply in characters is then mapped to the onset in this buffer to extract a snippet of appropriate whispering density that is further processed to deliver the reply sound event. Thereby longer replies sound more excited and faster without becoming unnatural.

Technically this method can again be regarded as a parameterized auditory icon mapping, but different from the approach in the chit-chat stream with discrete events in a sound font, we here realize a continuous selection process. While the actual psychophysical judgments of excitement may not increase strictly monotonously due to gaps and the details in the recorded whispering, the general trend will be dominating. The additional mappings are:

- sample file selection ← mood estimation, from :-) via :-| to :-(and nr. of '!' in the tweet.
- position in sample (degree of excitement) [begin, end] ← length of the tweet [0, 140 characters]
- The position and reverberation is consistent with the mappings for chit-chat events explained above.

Sound examples S7 contains a number of replies with increasing excitement (length of text). S8 contains a sequence of replies with average text length and different mood. They sound all neutral in space as they are the versions before any further post-processing.

5.3. The Hashtag sound stream

Hashtags are the parts of the tweets which we consider as relevant for judging the topics. Since hashtags can be freely invented by any user, it is impossible to set up a catalogue of possible strings and organize them in any meaningful way. As the sonification needs to create a sound in real-time without any intervention and reviewing by an editor, the sound needs to be synthesized from the string alone. Certainly, the first thought is to use any sort of text-to-speech system, or, to save time and avoid cluttering, to compress these spoken words just as *spearcons* do [7]. However, this would turn the sonification into a very verbal soundscape and possibly it would fail to convey what the Twitter dialogue is about. Thus we selected a more abstract way of encoding hashtags into sound-tags, oriented along two principles: (a) whenever a hashtag reoccurs, it has to be sonified by the identical sound as the previous one, (b) the hashtags cover a huge variety of sound events, just as words cover a huge variety of topics. Practically, we solve the problem by computing a hash which is reproducible for any hashtag string, with low risk that different strings result in the same hash value. We then use this hash to determine (i) a sound file in an extensible sample library with sounds from all areas of life, and (ii) details such as what snippet is extracted from the file and how it is distorted so that we obtain a very specific sound event for that hashtag. There is no easy way of generating a steady mapping between strings and sounds, so the hashtag #icad may sound very different from #icad2012. There is no underlying semantic analysis or categorization of words into classes such as economy, leisure, etc. Such extensions may be considered for specific continuations of the project.

Specifically the hashtag sound events are processed further using the following mappings:

- granular synthesis (sample, trigger rate, grain duration, etc.) ← hash(hashtag)
- sonority (how pitched vs. noise-like, via sample selection) ← ratio of consonants to hashtag length [0,1]
- delay, reverb, panning ← are consistent with chit-chat mappings.
- duration of hashtag events increase with decreasing global average activity (tweets per minute)

Perceptually, hashtag events stand out and appear as if in the foreground. Their unpredictability results in an element of surprise and should make listening to *tweetscapes* interesting even if there is no explicit interest to listen to it as a sonification. On longer and frequent listening to *tweetscapes*, users may remember and recognize certain sounds, such as #google, or #ff (short for #followfriday) on Fridays. Thematic changes are typically so slow that it is difficult to perceive them in continuous listening, but when listening to *tweetscapes* on different times or days, qualitative changes can be heard.

Sound example S8 and S9 are the hashtags for #papst (pope) and #piraten (a political party in Germany)[12]. Note that 'piraten' has more vocals and is somewhat more resonant. An example Tweetscape with these hashtags is discussed later on.

5.4. The Dominant Topics sound stream

As explained in Section 4, a ranking of hashtag frequencies is computed with a leaky integrator with 5 minute half-life. The technique to condense event streams into more complex events that represent aggregate properties was introduced in [8] and coined *Auditory Information buckets*. The idea is that a bucket collects information incrementally and flushes a more complex sound once the bucket is full. Here we take inspiration from this tipping bucket idea to define analogue structures that gather information about the dominance of topics. Only the three most filled collectors are selected for further sonification. Instead of a complex event localized

[12] as of Oct 2011, the algorithm has been refined meanwhile

in time we here create a continuous background sound that represents the hashtag sound as a stationary soundscape, so that the acoustic space is soaked with the idea of that topic. Certainly, this topic sound is the same as the corresponding hashtag, but using granular synthesis looped into a stationary pattern. Sound examples S10 and S11 present the corresponding topic sounds for the hashtags #papst and #piraten discussed in the previous section. To avoid a permanent overfilling of the sonic space with these topic sounds for the first 3 ranked topics, they are furthermore only added when they exceed a certain frequency (resp. counter value). The detailed mappings are:

- stereo panning [left, center, right] ← rank [2, 1, 3]
- level ← frequency counter [$f_{\min}$, $f_{\max}$], $-\infty$ below a threshold $f_{\min}$

5.5. Putting streams together

It is a difficult design task to tune all parameters and source sounds so that the individual streams work together as a coherent soundscape. Here particular effort was invested by the second author. The *tweetscapes* were first tuned according to our observation that the number of tweets rarely exceeded 5 per second using our filters. Sound example S12 is an example tweetscape with these data. However, modifications on the data interface to better capture the full German Twitter traffic led to an increase of the data volume per minute. In consequence a retuning was necessary since the soundscape became too densely filled. Sound examples S13 and S14 are two different versions for this more dense Twitter traffic. The solution to better cope with the available sonic space in time was to use the global activity (as already introduced above) to select the complexity and duration of events. This leads to less intrusive sounds once the intensity increases, as can be heard in sound example S14. From a sonification standpoint this procedure is debatable, since it breaks with the persistence of information. If we assume, however, that the main information lies in the level, frequency, echoes, reverberations and location, and we know that the density-driven selection process is reliable and reproducible we may simply adapt our listening habits and understand the soundscape correctly.

As a further extension we had considered including short verbal utterances that simply 'speak' a hashtag from time to time, at least on of the dominant topics. However, the speech synthesis lacked sufficient quality and robustness, given that hashtags are not necessarily words that can be spoken (e.g. #ff or #s21). So we canceled this path, yet it would probably be something valuable to consider for special application, such as for instance if visually impaired users showed an interest in using *tweetscapes*.

6. TWEETSCAPES VISUALIZATION

A frequent question that came from listeners who were first confronted with *tweetscapes* was 'what do the sounds actually mean?', 'what topics are discussed right now?'. We made clear that this is beyond the scope of the sonification and information we actively decided not to give. For the website at tweetscapes.de, fortunately the visual composer and 4th author Tarik Barri joined the team and created a real-time visual display (using his Versum [9]) that allows much better to connect the hashtag sounds with a particular meaning. The visual display shows the frontier line of Germany on a black background and dynamically creates colored light flashes at the location of the tweet. Furthermore, if it is a tweet with hashtag(s), the strings appear as text next to the light point. The synchronization of light and sound has two effects: (a) sound draws the attention to visual events, and (b) the textual display allows users to build up an association between hashtag sounds and their meaning. A particular feature is that replies to another tweet creates a visual arrow between the locations. This allows users to see how interconnected the Twitter space is.

7. TWEETSCAPES EXAMPLE SOUNDSCAPES

In this section we present three selected tweetscapes. The videos S15, S16, S17 are all captured from the live stream. S15 is a typical everyday activity. S16 represents a tweetscape at night – this is a much less populated soundscape. Finally S17 is a Tweetscape at a specific event. More detailed explanation will be given on the website with the sound examples. Our general experience is that the visual part is quite absorbing and draws the attention very much. So we recommend listening to the tweetscapes also with closed eyes, to investigate whether you can differentiate the situations by listening, or recognize or identify repeated topics.

8. EMBEDDING TWEETSCAPES INTO THE RADIO PROGRAMME

Tweetscapes was tailored to a particular role within Deutschlandradio Kultur's radio drama, documentary and sound art program: In this department, productions rarely match the precise length of their respective slots. The resulting time gaps are usually filled with generic music to be faded out when the news come in. In order to artistically shape this gap, Deutschlandradio Kultur's former sound art editor Götz Naleppa introduced a special format in 1998: 'Das Geräusch der Monats' (the noise of the month) were 5 minute sound art compositions designed to be faded in and out at any given time. This format was replaced by the Sonarisations in October 2011.

The piloting *tweetscapes* project meets the challenges of this particular slot in many ways:

- since *tweetscapes* taps into a live data stream, it can be faded in and out at ease
- *Tweetscapes* presents an artistic take on a topic of general interest
- the elaborate sound design makes *tweetscapes* equally accessible as a musical composition for a larger public and as a carrier of relevant information for experts

Embedding *tweetscapes* into the structures of Deutschlandradio Kultur required a number of thorough preparations. First, the concept needed to be communicated within the hierarchy and different departments concerned. The risks of real-time rendition, with unpredictable sound output needed to be tackled, both in terms of reliability (i. e. what if the synthesis fails?) and quality (i. e. what if Twitter traffic develops so that the tweetscape is unacceptable?). Furthermore, the embedding demanded significant technical infrastructure, from setting up a dedicated computer with the high security standards inside the intranet of the broadcasting station, to procedures to backup and access for maintenance.

Once these steps were taken, the integration into daily use required the production of programs explaining the purpose and idea of the project, as well as the setup of a project website, the edition of short texts for moderators to read before tweetscapes are played, etc.

Finally, the relaunch of tweetscapes.de with the audiovisual live stream challenged the means of a public broadcaster in terms of supporting online projects. However, the website and visualization have proven perfectly complementary to the sound stream, offering greater transparency and accessibility for a wide range of users.

9. DISCUSSION

With *tweetscapes*, we have – for the first time – established sonification into the regular program of a national broadcasting station. This project allowed us to learn many lessons on many levels. One level is the interdisciplinary communication: drawing together radio professionals, sonification researchers and artists/composers proved to be highly beneficial both for the involved persons that appreciated the different views and for the project since it offered to go beyond typical paths that probably would have been taken if not the mutual negotiations helped us to find a view 'in between' the poles. Our take is that it is definitively worth the effort.

The second level is the one of sonification for public media: we were surprised by the huge interest from media and press to report about *tweetscapes*, in fact the project launch event was highly visible due to press releases from DPA and even made it to several nation-wide newspapers. The reception of the project, however, showed a wide range of comments, from 'useful' / 'nice artwork' to 'waste of time'. Only few recognized *tweetscapes* as an example of sonification and understood the idea behind it, which is the general idea to represent complex information reliably by using non-speech sound. They related to *tweetscapes* more as 'making music from Twitter'. Mostly the question arose 'What is the practical use of listening to *tweetscapes*?' Indeed, the practical use is very limited – it is the *idea* that we here wished to transport. Understanding the Twitter space as such by listening is a new experience and that may or may not be inspiring for the listener. When getting in contact with public media, apparently there is the need and tendency to break complex ideas down into the most basic and raw concepts that anybody can connect with. This led to headlines such as 'turning Twitter into music', a phrase where sonification researchers will probably disagree.

9.1. Interactive participatory radio-making

On another level we see the potential of *tweetscapes* to establish something really new in radio broadcasting: the ability that radio listeners can via *tweetscapes* participate and influence the radio broadcast in real-time. This may on first sight only appear to be a neat gimmick, yet on second sight, it may allow completely new forms of radio shows. For instance, imagine that the moderator can ask the audience what they find most interesting to focus on – the radio listeners in turn tweet their opinion using pre-determined hashtags, and they can experience in real-time the distribution and frequency of opinions of others. The moderator can then use this information to refine or adapt the program or to select the next questions in an interview, etc. *Tweetscapes* thus provides not only a new 'unconventional view' on Twitter, it opens and suggests new forms of interactions in radio culture.

10. CONCLUSION

We have introduced *tweetscapes*, a real-time sonification system that allows users to become aware of Twitter traffic by listening. We have reported the goals, design ideas, methods, sonification streams, and played concrete examples for the various elements in *tweetscapes*. The multi-stream event-based sonification uses established parameter-mapping techniques and less frequently used ideas such as sound fonts and continuous sample selection for parameterized auditory icons. We explained how *tweetscapes* has been integrated into the regular program of Deutschlandradio Kultur and we have shown an audio-visual extension (live stream) which is featured on the project website. Finally we outlined some new ideas of how *tweetscapes* could in future inspire new forms of participatory interactive radio. *Tweetscapes* is the pilot project for the continued series 'Sonarisations' that

aims at making sonification publicly known by featuring its possibilities in a nation-wide radio program.

11. ACKNOWLEDGMENT

We thank Deutschlandradio Kultur who enabled the realization of *#tweetscapes*. We thank the German Research Foundation (DFG) and the Center of Excellence 277 Cognitive Interaction Technology (CITEC) that enabled this work within the German Excellence Initiative. We thank Sam Auinger, Holger Schulze, Martin Supper and Georg Spehr for early discussions that lead to *#tweetscapes*. We thank the staff at Twitter for their help, namely Katie Jacobs Stanton, Jason Costa und Carolina Janssen.

12. REFERENCES

[1] P. Vickers, "Sonification for process monitoring," in *The Sonification Handbook*, T. Hermann, A. Hunt, and J. G. Neuhoff, Eds. Berlin, Germany: Logos Publishing House, 2011, ch. 18, pp. 455–491. [Online]. Available: http://sonification.de/handbook/chapters/chapter18/

[2] A. D. N. Edwards, "Auditory display in assistive technology," in *The Sonification Handbook*, T. Hermann, A. Hunt, and J. G. Neuhoff, Eds. Berlin, Germany: Logos Publishing House, 2011, ch. 17, pp. 431–453. [Online]. Available: http://sonification.de/handbook/chapters/chapter17/

[3] T. Hermann, J. M. Drees, and H. Ritter, "Broadcasting auditory weather reports – a pilot project," in *Proceedings of the International Conference on Auditory Display (ICAD 2003)*, E. Brazil and B. Shinn-Cunningham, Eds., International Community for Auditory Display (ICAD). Boston, MA, USA: Boston University Publications Production Department, 07 2003, pp. 208–211.

[4] H. Schulze, "Sonarisationen. ein projekt künstlerischer forschung des deutschlandaradio kultur berlin," in *Das geschulte Ohr*, ser. Sound Studies. Bielefeld, Germany: transcript Verlag, 2012, vol. 4, pp. 283–298.

[5] T. Hermann, "Taxonomy and definitions for sonification and auditory display," in *Proc. 14th Int. Conf. Auditory Display (ICAD 2008)*, B. Katz, Ed., ICAD. Paris, France: ICAD, 06 2008.

[6] W. W. Gaver, "Using and creating auditory icons," in *Auditory Display*, G. Kramer, Ed., ICAD. Reading, MA: Addison-Wesley, 1994, pp. 417–446.

[7] B. N. Walker, A. Nance, and J. Lindsay, "Spearcons: speech-based earcons improve navigation performance in auditory menus," in *Proc. Int. Conf. Auditory Display (ICAD 2006)*, T. S. et al., Ed., ICAD. London, UK: Department of Computer Science, QMC, University of London, 2006, pp. 63–68.

[8] T. Hermann, M. H. Hansen, and H. Ritter, "Sonification of markov-chain monte carlo simulations," in *Proceedings of 7th International Conference on Auditory Display*, J. Hiipakka, N. Zacharov, and T. Takala, Eds., ICAD. Helsinki University of Technology: Laboratory of Acoustics and Audio Signal Processing and the Telecommunications Software and Multimedia Laboratory, 07 2001, pp. 208–216.

[9] T. Barri, "Versum: audiovisual composing in 3D," in *Proc. 15th Int. Conf. Auditory Display (ICAD 2009)*, Copenhagen, Denmark, 06 2009.

A MODULAR COMPUTER VISION SONIFICATION MODEL FOR THE VISUALLY IMPAIRED

Michael Banf & Volker Blanz

Universität Siegen,
Media Systems Group, Germany
{banf, blanz}@informatik.uni-siegen.de

ABSTRACT

This paper presents a *Modular Computer Vision Sonification Model* which is a general framework for acquisition, exploration and sonification of visual information to support visually impaired people. The model exploits techniques from Computer Vision and aims to convey as much information as possible about the image to the user, including color, edges and what we refer to as *Orientation maps* and *Micro-Textures*. We deliberatively focus on low level features to provide a very general image analysis tool. Our sonification approach relies on MIDI using "real-world" instead of synthetic instruments. The goal is to provide direct perceptual access to images or environments actively and in real time. Our system is already in use, at an experimental stage, at a local residential school, helping congenital blind children develop various cognitive abilities such as geometric understanding and spatial sense as well as offering an intuitive approach to colors and textures.

1. INTRODUCTION

Sonification is defined in [1] as "use of non-speech audio to convey information". It describes both a scientific and an artistic discipline. In recent times it is even used to audibly browse the RNA structure [2] or model an acoustical representation of the standard model of particle physics [3]. In this paper we address the sub-field of developing sonification methods to support visually impaired people. In the last years, we have seen a number of special purpose devices that help visually impaired people to solve everyday problems, such as finding their way [4] or reading text. These devices tend to be restricted to a specific problem by extracting very specific information from the input data or by relying on additional information, such as positioning systems. In contrast, our goal is to develop a general-purpose system that can be used in a wide range of situations. Our system analyzes visual data using general image descriptors from computer vision, and maps these to auditory signals. Unlike approaches that use, for example, face detectors, we deliberately focus on low-level descriptors such as colors and edges in order to obtain a device that can sonify any given image and help to solve any given question that persons with normal vision could solve: We want to enable users to recognize what kind of scene is shown, find objects in this scene, or explore the photos of a friend. Unlike high-level image analysis, our device gives direct feedback on what is where in the image. We are inspired by the way how visually impaired persons can explore reliefs of scenes with their fingers and get a direct haptic experience of the shapes and locations of objects – they can feel what the peak of a mountain or a what cloudy sky is. Due to the simplicity and directness of the sensory mapping from visual to auditory, we harness the human ability to learn, so we consider the brain of the user as part of the system.

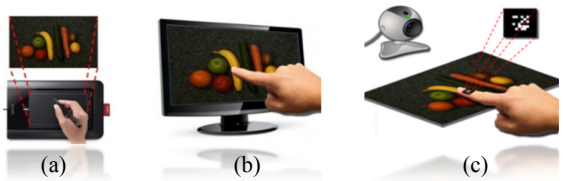


Figure 1: Several exploring interfaces: (a) Pen to tablet interface with absolute coordinates. (b) Touch screen working directly on the acquired image. (c) Finger position *(x,y)* is tracked using a marker system and estimated within the image

There has been done previous work on the sonification of low-level characteristics of images for visual impaired. [5] generates sounds depending on a pixel's lightness and its position within the image. [6] demonstrates a technique, using color patches within images as chromatic patterns that can be put together to form a melody. [7] uses color attributes to filter an underlying white noise using *Subtractive Synthesis*. [8] mixes 8 pre-recorded musical timbres depending on the quantity of 8 hue values within an image. All such synthesized sounds are not easily interpreted at first. Especially the last two approaches map all attributes of a certain color onto a single sound parameter, such as timbre, pitch or frequency spectrum. Such mappings are hardly reversible and might lead to identical sound synthesis from different input color combinations. Our contribution to previous work is to present a new and holistic approach to color and texture sonification, which is intuitive enough not only to be understood and applied by congenital blind, but also helps to convey e.g. the concept of colors and color mixing itself. We consider our approach to be holistic, as it maps each attribute of a particular color within a color-space – such as hue, saturation and lightness - to an intuitive, but separate, counterpart within the sound space at once. Additionally we map texture features to unique sound in the same way. The sonification of colors is important for two reasons, first, to offer a way to congenital blind people to understand colors and to be able to communicate with non-visually impaired about such fundamental quality of human vision. Second, in illuminated environments, colors are crucial in the process of detecting objects. If our sonification is intuitive enough, a person will quickly learn to understand the concept of colors and textures as well as the audification, recognize objects, interpret images and develop their own strategies. Designing such a system, we face the following challenges:

- Sensory input: We propose a system that analyzes images that users may find in a photo collection, on the internet, or capture with a still camera. Future devices could allow for depth information (stereo or time-of-flight devices), motion or other input.

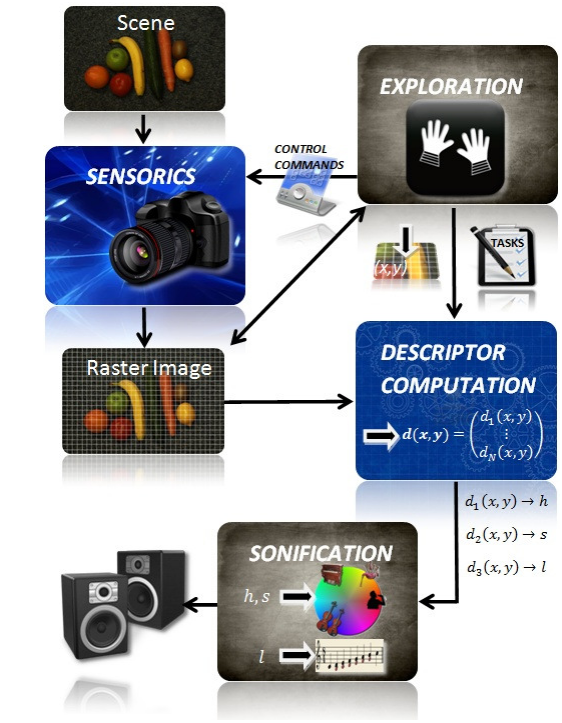


Figure 2: The *Modular Computer Vision Sonification Model.*

- Design the most appropriate visual descriptors to represent particular image values. They should be informative, general and stable under transformations such as illumination and pose.
- Define a way to sonify these image descriptors. The goal is to convey as much information as possible without interaction between the channels as well as to give an auditory perception that enables users to develop an "intuition" about the visual data.
- Develop an exploration paradigm. Human vision has many aspects of parallel processing: Much of the visual pathway in transmits information from different parts of the visual field in parallel, and pre-attentive vision (pop-out effects) indicates parallel processing on higher levels. In contrast, an auditory signal mostly is a sequential data stream. This implies that it is hard to map an entire image to a single, constant auditory signal. Therefore, we decided that users should explore the image locally, for example on a touch screen.

Our paper is on the interface between computer vision and sonification. Our contribution is the general concept of a *Modular Computer Vision Sonification Model* with the components *Sensorics – Exploration - Machine Vision – Sonification – Human Learning*, and a specific setup that implements this concept and that we present and evaluate below. As the notation "*Computer Vision*" implies, we focus upon working with visual data. More abstract models for the process of data sonification in general have been formulized e.g. in [9]. The importance of interaction in sonification is argued in [10].

2. THE MODULAR SONIFICATION MODEL

Figure 2 gives an overview of the general concept of our *Modular Computer Vision Sonification Model.* Stage one is the

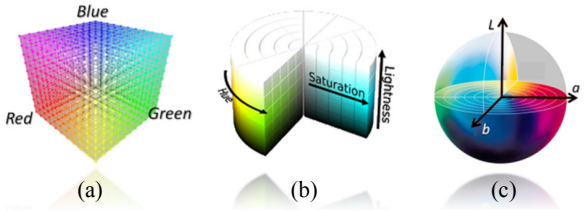


Figure 3: Color Models: (a) *RGB.* (b) *HSL.* (c) *CieLab*

acquisition of both a rasterized image ***I*** as well as a particular pixel position *(x,y)* where the image is to be explored. Further, several tasks about what features shall be calculated and sonified are determined and passed through during exploration. Next step in the process is the creation of a *pixel descriptor* ***d****(x,y)* – a vector gathering all the information to be sonified. The descriptor captures image information in the local neighborhood of *(x,y)* and is calculated using computer vision techniques. This information might be very fundamental such as color, texture, edges as well as more complex such as an estimation of depth or whether the current pixel belongs to a face. Every feature *i* makes an element $d_i(x,y)$ of the pixel descriptor ***d****(x,y)* and is transferred to the sonification unit. In the next sections, we discuss each module and describe our specific design and its implementation.

3. SENSORICS

The sensory module acquires the data to be sonified. In the system presented in this paper, we rely on still images that are available as files. It is easy to apply our system to still images from a camera operated by the user, or on images from web pages. Future extensions may include depth information (from stereo vision or depth sensors), infrared images, GPS positioning, and motion information in videos.

4. EXPLORATION

Our exploration module has two major tasks. First, during the image acquisition, it allows the user to transmit control commands to the sensory module such as activating certain sensors or controlling their viewing direction. Second, it enables the user to navigate within an image, passing on its position *(x,y)* to the descriptor computation unit, along with several tasks that determine which features of ***d****(x,y)* shall be sonified at all.

4.1. Navigation

Navigating within an image requires an appropriate interface. The computer-mouse, which is popular among users with normal vision, drops out as it does not deliver any absolute coordinates, which are necessary for a blind user to know the position in the image. Hence, we worked with several interfaces, as shown in Fig. 1, to see what suits best to a blind person. The pen – tablet interaction method functioned far better than the mouse, as it can be set to absolute coordinates. However, it turned out, that a direct touch helps to orient within the flat image, as analogous to moving the tip of the finger along a relief. Touch pads without a pen usually provide only relative positioning (similar to a mouse). For training and several user studies in section 7 we utilized a touch screen that allows the user to interact direct with the image plane (without seeing the image). Even though it does not make sense for a

(a) (b)

Figure 4: (a) Raster Images. (b) *Bilateral filtered* images.

blind person to buy a touch screen, the absolute positioning proved very successful. Therefore we implemented a more effective and far cheaper interaction method, applying a camera-based finger position tracking based on *ARToolKitPlus* pose tracking system [11]. The same camera as to acquire the visual data is utilized to detect a marker, attached to the user's fingernail, which is thereafter calculated back to estimate the fingers position within the image.

4.2. Control Gestures

The use of a pose tracking system turns out to be of great interest in controlling the whole system as well. As the utilized *ARToolKitPlus* is able to deal with several markers at once it allows us to recognize many finger and hand gestures to submit control commands to the sensory module as well as sonification tasks to the descriptor computation unit. In contrast, a keyboard or virtual buttons on the touch screen are difficult to operate for blind users.

5. PIXEL DESCRIPTOR COMPUTATION

The *pixel descriptor* ***d****(x,y)* holds all relevant features to be sonified at an image position *(x,y)*. We focus on fundamental characteristics such as colors and what we call *Orientation maps* and *Micro-Textures*. However, as intended, the module can be extended to extract and store more complex information.

5.1. Color Information

There are different color systems with several motivational backgrounds, as shown in Fig. 3. The *RGB* model uses additive mixtures of red, green and blue. It is motivated by the human eye receptors [12] and applied, e.g. in many display devices. However, providing a non-visual access to colors, as in our case, requires a more intuitive system, especially for congenital blind persons. This is why we prefer the *HSL* model [13], where each color value is described by hue *h*, saturation *s* and lightness *l*. What makes color sonification difficult is the fact that color values often change rapidly from pixel to pixel even if there are only minute variations in textures and materials. Often, the reason is image noise by the camera. It is obvious that such changes clearly overburden a blind user. Therefore we smooth the image patch around the pixel position *(x,y)* based on *Bilateral Filtering* [14], as shown in Fig. 4, which filters noise while preserving edges within an image and will play a significant role in finding orientation maps. Subsequently, we use the smoothed color values as the first three elements of our pixel descriptor ***d****(x,y):*

$$d_1(x,y) = h_{smooth}(x,y), \quad d_2(x,y) = s_{smooth}(x,y), \quad d_3(x,y) = l_{smooth}(x,y)$$

5.2. Orientation Maps

The rationale behind *Orientation Maps* and *Micro-Textures* is

(a) (b)

Figure 5: (a) *Gaussian Pyramids* on red, green and blue channels. (b) *Gabor Transform* response images.

to create something like an acoustical relief that allows the user to hear what is under his fingers, instead of feeling it. While micro-textures - explained in the next section - express the overall roughness characteristics of a particular patch, orientation maps represent dominant structures within the image. We consider dominant structures single or repetitive sets of significant edges of the same orientation and a particular direction of propagation. Our method is based on the observation that standard edge detectors such as *Canny* [15] produce multiple edges and spurious, misleading signals that confuse the user. Therefore the calculation of orientation maps involves filtering important from distracting structures, which may be motivated biologically from the *Surround Inhibition* in the human visual system that improves contour detection [12]. Moreover, regular repeating patterns of a certain size tend to be human made, unlike the more fractal patterns that are often found in nature [16]. Finding human made structures is important when using the system to orient within an environment to find windows, doors, ways, tables, shelves and so forth.

5.2.1. Calculating Orientation Maps

In the first step of calculating orientation maps, we use cascades of *Median* [17] and *Bilateral Filtering* to suppress both noise and small corners, as shown in Fig. 4 (b). Then, we reduce the spatial resolution of the red, green and blue color channels to obtain a *Gaussian Image Pyramid* [18], as shown in Fig. 5 (a). On each level, the width and height is reduced by a factor of 2. This reduction process removes low-scale variations which may be irrelevant for the task of the user. By later combining information of different layers using the image pyramid we can select the most appropriate resolution for each visual feature. Next, we perform a *Gabor Transform* [19] on each channel separately and build a final response image by measuring all individual responses. The transform relies on *Gabor Wavelets* $\psi_{\varphi,\nu}(x,y)$ [20] of the form:

$$\psi_{\varphi,\nu}(x,y) = \frac{\|k_{\varphi,\nu}\|^2}{\sigma^2} e^{-\frac{\|k_{\varphi,\nu}\|^2 \|(x,y)\|^2}{2\sigma^2}} \left[e^{ik_{\varphi,\nu}(x,y)} - e^{-\frac{\sigma^2}{2}} \right] \quad (1)$$

where the parameters φ and ν define the orientation and scale of the Gabor kernel and σ is the standard deviation of the Gaussian window in the kernel, i.e. the size of the window. $k_{\varphi,\nu}$ is the wave vector, combining orientations and the spatial frequency in the frequency domain. Gabor Wavelets are widely used in computer vision [20], because they provide an analysis of spatial frequency that is local, unlike the global analysis in a Fourier Transform [17]. To filter orientation maps, we choose $\nu = 1$ and later combine their particular response – inspired by the cascading of several *simple cells* to form *Complex Cells* [12] in the human visual cortex. We apply them in 32 orientations $\varphi = \{0°, 5.625°, 11.25°, 16{,}875, \ldots, 180°\}$. For the carrot and the orange in Fig. 4 (a), such response images are shown in Fig. 5 (b). The 32 orientations φ are visualized by gray scale values. We

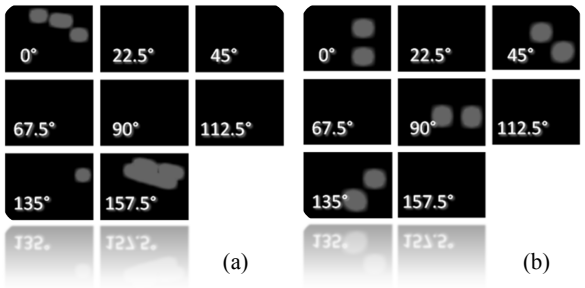


Figure 6: *8 Orientation Maps* for (a) carrot and (b) orange image

quantize the response image so that the each of the 32 orientations φ is mapped to the next of 8 orientations $\theta = \{0°, 22.5°, 45°, 67.5°, 90°, 112.5°, 135°, 157.5°\}$. Each θ is represented in an individual gray scale image $\boldsymbol{G}_\theta$, where $\boldsymbol{G}_\theta(x,y) = 255$, in case of an edge and $\boldsymbol{G}_\theta(x,y) = 0$ otherwise. One fundamental idea of orientation maps is that the user should not have to follow contours of objects or structures to estimate their silhouette, which would be tedious and slow. To make it easier for users to find contours, we distribute them around the edge by a kind of diffusion approach. Therefore, we calculate for each image position (x,y) the *variance* $\sigma^2(x,y)$ of values $\boldsymbol{G}_\theta$ on its local neighborhood on each of the 8 images $\boldsymbol{G}_\theta$, to obtain what we call *Orientation Maps* $\boldsymbol{O}_\theta$:

$$O_\theta(x,y) = \sigma^2(x,y) = \sum_{i=-w/2}^{w/2} \sum_{j=-w/2}^{w/2} (G_\theta(x+i, y+j) - \mu(x,y))^2$$

with

$$\mu(x,y) = \frac{1}{w^2} \sum_{i=-w/2}^{w/2} \sum_{j=-w/2}^{w/2} G_\theta(x+i, y+j) \qquad (2)$$

where $\mu(x,y)$ *is* the mean and w the size of the local neighborhood. *Orientation Maps* $\boldsymbol{O}_\theta$ for the carrot and orange (Fig. 4 (a)) are shown in Fig. 6. Each coherent patch of gray scale pixels, on each orientation map is referred to as *Orientation Patch* $V_{\theta,i}$. As the diffusion approach might cause overlap in image positions (x,y) of different oriented orientation patches, as illustrated in Fig. 7 (a), we now compare orientation patches and emphasize the dominating ones while suppressing insignificant and therefore distracting others.

5.2.2. A Topological Representation of Orientation Patches

We consider orientation patches as dominant if they have a certain size – which is the number of their pixel positions. So far, we handle 4 cases. Case 1: If an image area is dominated by two very big orientation patches of almost equal sizes, both above a certain threshold t_{size} and such patches differ in orientation by more than 22.5°, they are retained as coexisting. This happens to be the case for a rectangular grid or a wall of bricks, where two orthogonal orientations are permanently present and form the particular textures of the image region. Case 2: If the image area is dominated by two orientation patches, both greater than t_{size} , having an orientation difference of only 22.5°, which is the smallest possible difference, these patches are combined into a single orientation patch by merging the smaller one into the bigger one. Each pixel of the smaller orientation patch is assigned to the orientation map of the bigger one. After that, the smaller orientation patch is erased from its orientation map. Case 3: If the image area contains a large orientation patch, with a size greater than t_{size}, and further patches, whose sizes are below $t_{size,}$ such smaller patches are

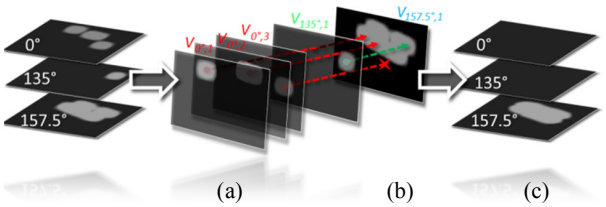


Figure 7: (a) Split initial Orientation Maps. (b) Arrows: centers of $V_{0°,1}$, $V_{0°,2}$ and $V_{3,1}$ lie within $V_{2,1}$. Cross: center of $V_{1,3}$ does not lie within $V_{2,1}$. (c) Final Orientation Maps.

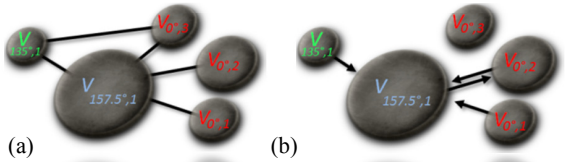


Figure 8: (a) G_o of Fig 6 (a). (b) G_c of Fig. 6 (a).

either (a) merged into the big one, as described in case 2, or (b) deleted from their orientation maps, depending on whether their particular centers lie within the large orientation patch or not. Case 4: If the image area contains several orientation patches, whose sizes are below t_{size}, they are merged as in case 2, in case both their center positions overlap and their orientation difference is equal to 22.5°. Otherwise they coexist. To implement the rules described in this section, we build a topological representation of all overlapping orientation patches in that area by the following procedure. At first, we apply a *contour finding algorithm* [21] on each orientation map $\boldsymbol{O}_\theta$ that retrieves a sequence of all contour pixels as well as all enclosed image positions found within the map. Each contour, as well as the enclosed pixel positions, belong to an orientation patch $V_{\theta,i}$. We now represent each $V_{\theta,i}$ as a single image, as illustrated in Fig. 7 (a). So far, for each $V_{\theta,i}$ we have its size and can calculate its center of gravity. Starting with a particular orientation patch $V_{\theta,i}$, we now check pixel by pixel for overlaps with each different oriented orientation patch. In case of overlapping, we compute the number of mutual pixel positions and whether the center of one orientation patch lies within the other. The results can be processed by *graph theory* [22] in the following way. Overlapping orientation patches can be modeled as an *undirected Graph* $G_o = \{V,E\}$, where *knots* V represent all orientation patches and *edges* E represent the existence of an overlap between a pair of different oriented orientation patches. A second graph G_c is set up to represent only such overlaps, where at least the center c of one of the two orientation patches involved is found inside the related orientation patch, as visualized in Fig. 7 (b). In this case connections may be only in one direction, so G_c is a *directed graph*. Both graphs G_o and G_c, for all orientation patches of the carrot image (Fig. 6 (a)), are illustrated in Fig. 8. Such topological representations can now be used to find which of the 4 previously mentioned cases fit. For the carrot image we find, based on G_o and G_c (Fig. 8 (c)) that $V_{0°,1}$ and $V_{0°,2}$ merge into $V_{157.5°,1}$ applying case 3a and $V_{0°,3}$ is deleted, applying case 3b. Hence, the final orientation maps $\boldsymbol{O}_{0°}$, $\boldsymbol{O}_{135°}$ and $\boldsymbol{O}_{157.5°}$ are shown in Fig. 7 (c). In contrast, the Graph G_o based on the orientation maps of the orange image, Fig. 6 (b), would result in a ring shaped connected structure, where every knot would be about the same size. Hence, such orientation patches are left as they are, according to case 4. Eventually, we can assign the shares of all eight orientation maps at a particular pixel position (x,y) to the pixel descriptor

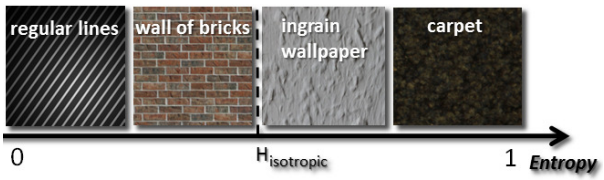


Figure 9: 4 types of textures with increasing *Entropy*.

$\boldsymbol{d}(x,y)$. We define $\Omega_\theta(x,y) \in \{0,1\}$ to be either 1, in case $\boldsymbol{O}_\theta(x,y) > 0$, or 0, if $\boldsymbol{O}_\theta(x,y) = 0$.

$d_4(x,y) = \Omega_{0°}(x,y)$, $d_5(x,y) = \Omega_{22.5°}(x,y)$ $d_6(x,y) = \Omega_{45°}(x,y)$,
$d_7(x,y) = \Omega_{67.5°}(x,y)$, $d_8(x,y) = \Omega_{90°}(x,y)$, $d_9(x,y) = \Omega_{112.5°}(x,y)$,
$d_{10}(x,y) = \Omega_{135°}(x,y)$, $d_{11}(x,y) = \Omega_{157.5°}(x,y)$

Note that as borders of orientation patches fade out, caused by our *variance* σ^2 *diffusion* approach, we are also able to have $\Omega_\theta(x,y)$ run from 0 to 1 in several steps: $\Omega_\theta(x,y) \in \{0,\ldots,1\}$.

5.3. Micro Textures

What we call *Micro-Textures* in our system captures the roughness and local structure at a point in an image. Examples of textures that we want to distinguish are shown in Fig. 9: regular lines, brick walls, ingrain wallpaper or carpet. In the Computer vision literature, there are many approaches to describe texture, even though it is difficult to give a general definition. [13] describes three different groups of texture measures. First, there are *First Order Statistical Texture Measures* [17] such as the *mean* μ and *variance* σ^2 of color values in a local neighborhood. Second, there are *Second Order Statistical Texture Measures*. These texture measures do not analyze single intensities, but correlations between pairs of pixel values. An example of this approach are *Haralick texture measures* based on *Co-Occurrence-Matrices* [23]. Unfortunately, a major drawback are high calculation costs. There have been efforts to speed up the processing using GPU [24]. Finally there is *Spectral Image Analysis* such as e.g. the *Fast Fourier Transform* [17] or *Gabor Transform*. To build micro-textures we use the *Gabor Transform* and compute the *Entropy* as a texture measure.

5.3.1. Entropy as Texture Measure

The *Gabor Transform* is applied to the original image and not to the bilateral filtered image, as we now want to preserve roughness information. *Entropy*, generally measures the disorder within a physical system, and is used in various scientific fields. Having zero entropy means to have maximum information about the state of a system [25]. In information theory, it is formulated as:

$$H = -\sum_{i=1}^{N} p_i \log p_i \quad \text{with} \quad p_i = \frac{N_i}{N} \qquad (3)$$

We calculate the *Entropy* $H(x,y)$ for each pixel position (x,y), based on the *Gabor Transform* response within a local neighborhood. The variable p_i is the probability for a certain orientation φ_i estimated from its occurrence N_i divided by the total number N of all orientations φ that occur within the window. Based on $H(x,y)$ we can now measure the roughness of an image region and assign it to the last element of $\boldsymbol{d}(x,y)$:

$$d_{12}(x,y) = \begin{cases} 0 \ (\textit{smooth surface}), & \text{if } N < N_{minimum} \\ 1 \ (\textit{anisotropic roughness}), & \text{if } 0 \le H(x,y) < H_{isotropic} \\ 2 \ (\textit{isotropic roughness}), & \text{if } H_{isotropic} \le H(x,y) \end{cases}$$

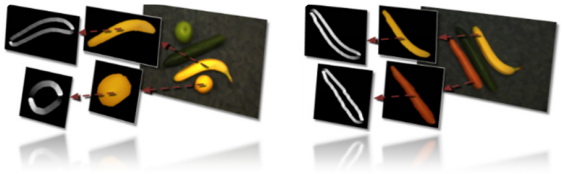


Figure 10: Some results of color segmentation.

5.4. Grabbing Objects

We utilize state of the art segmentation algorithms to separate the image into regions that are likely to show different real life objects. The goal is to help users find and scrutinize objects and other entities in the image, based on orientation maps and micro-textures calculated for such parts only. Image Segmentation or more precisely a foreground / background segmentation - is performed using *Gaussian Mixture Models* calculated using *Expectation-Maximization* and *Graph Cuts* [26], [27], [28]. First, the user moves over an area of interest and initiates the segmentation procedure by pressing a button or by gesture. Based on the users current position (x,y) we apply a *flood-fill* algorithm [29] that iteratively adds pixels to an area around (x,y), if their color distance to the average color of the region is below a threshold. The color distances are calculated as the Euclidean distance $\|\ldots\|$ in *CieLab* [13] Color Space, as illustrated in Fig. 4 (c). All these selected pixels are marked as "definite foreground", all others as "probably background", and both groups serve as first segmentation estimation and as input to the *Expectation Maximization* and *Graph Cut* algorithms, which then calculate the final segmentation. Fig. 10 shows some exemplary results of the segmentation, as well as the responses of the *Gabor Transform* applied to such segmentations. The whole segmentation process takes approximately *4.5* seconds and can therefore be considered for interactive usage.

6. SONIFICATION CONCEPT

The previous sections dealt with extracting features to form the pixel descriptor $\boldsymbol{d}(x,y)$. We now discuss the question of how to sonify those features. A great challenge is to avoid conflicting signals and information overload, as well as the transformation of quasi-static 2D image data into a dynamic audio stream. Though humans can distinguish many attributes such as *pitch, volume, ADSR-Curve, timbre, roughness or vibrato*, it is still impossible to transport all potential descriptors of visual information simultaneously. Unlike approaches that sonify a whole image sequentially e.g. by scanning its pixels row by row [30] we want the user, as already described, to fully interact with the visual data in real- time and to be able to hear what is currently under his finger. Second, we want a method to simultaneously sonify features such as color, orientation maps and micro-textures and even more, instead of focusing on a single feature such as the progression of edges [31], [32]. Third the sonification model should meet aesthetical demands that are important for comfortable and extensive usage.

6.1. Sonification of Color Information

Sonification systems may use different techniques of sound synthesis, such *as Subtractive Synthesis*, *Additive Synthesis, Granular Synthesis, Physical Modeling* or *FM Synthesis* [33]. The methods presented in this paper rely on sounds from common instruments based on the General MIDI (GM) Standard (based on *Wavetable Synthesis* [33]). Visual impaired people

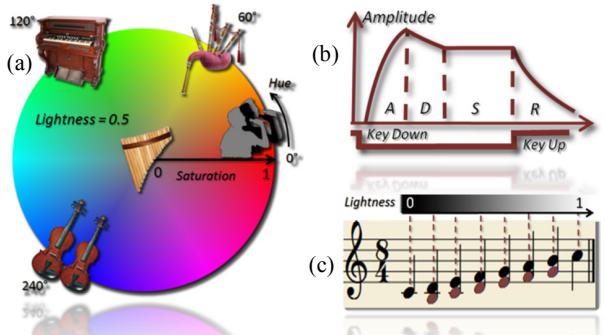
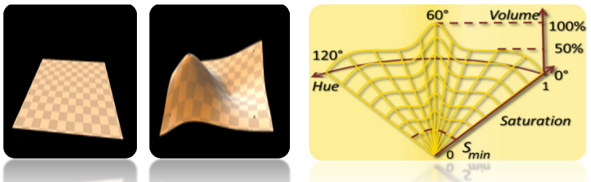


Figure 11: (a) *Complementary Instruments* represent pairs of *opponent colors*. (b) *ADSR* Envelope for the hit of a piano key. (c) Lightness $l \in \{0,\ldots,1\}$ and musical scale each l is assigned to. Brown notes are the added thirds.

may find this a comfortable and soon a familiar way to get perceptual access to colors and textures. Instead of strictly learning particular associations between instruments and colors (which they do not see and therefore have to memorize) we want to help them to build connections between sonification signals and objects of their daily life. In fact, in our experiments, we often heard participants say that an image region "sounds like a tomato" or any other object rather than announcing the correct mixture of colors, such as red – as in case of the tomato.

6.1.1. Complementary Instruments

As we use common instruments to sonify visual information, we propose a concept that represents each color value in the *HSL* model as a mixture of instruments, inspired by Hering's *theory of opponent colors* [12]. In principle, we use what we call *Complementary Instruments* to represent the *opponent color pairs* red-green and blue-yellow, as shown in Fig. 11 (a), and later combine adjacent instruments to represent color mixtures. As no mixture of a pair of opponent colors exists [12], there will be no mixture of a pair of complementary instruments in the sonification model either. Further we apply a musical scale to represent the luminance scale from black to white. Complementary instruments therefore must guarantee certain characteristics. First, they must possess a relatively stable frequency spectrum over time. That means that in terms of *Attack-Decay-Sustain-Release - Amplitude envelope (ADSR)*, as shown in Fig. 11 (b), they should have a short *Attack-* and *Decay-*, an infinite *Sustain-* and a short *Release-Phase*. To avoid mutual masking of instruments, their frequency spectra should have narrow bandwidths (i.e. little noise components). In addition to appropriate *ADSR*-characteristics, there are further criteria that a set of 4 complementary instruments has to fulfill: *Separability* ensures that instruments, assigned to adjacent colors can be clearly distinguished even when they are played as mixtures. This criterion does not need to be met by complementary instruments. Second we need *Uniqueness:* Even complementary instruments need to be unique enough to be associated with its particular color. Finally, we want to make sure that mixtures of instruments do not sound like other, new instruments. Fig. 11 (a) shows our final selection of instruments: Choir (red), bagpipe (yellow), organ (green), strings (blue) and flute (white, black, gray). The software allows users to assign own selection of preferred instruments. The specific role of gray-scale, black and white with only one instrument will be explained in the next section.



(a) (b) (c)

Figure 12: (a) 2D spline. (b) Same 2D spline deformed given 6 control values c. (c)*Volume Shape* $\vartheta(h,s)$ scheme around yellow.

6.1.2. The Concept of HSL-Color Sonification

As explained in section 5 the *HSL* color model describes a certain color using hue h, as an angle from 0° to 360°, lightness l and saturation s. This color information of a pixel is stored in the first three element of $\mathbf{d}(x,y)$: $h = d_1(x,y)$, $s = d_2(x,y)$, $l = d_3(x,y)$. Based on our idea to assign complementary instruments to certain hues, we sonify intermediate color tones as mixtures of two adjacent instruments, and represent the color mixture ratio by their partial volume. The fade of saturation s, moving inward to the center of Fig. 11 (a), is considered as a general absolute decrease in volumes of any two color instruments playing simultaneously, while their relative volume ration is maintained. However, below a certain threshold s_{min} we regard the color as gray and sonify it using a single instrument, the flute. In general, gray is not considered a color, and the *HSL* model assigns it an arbitrary hue h = -1 and a saturation s = 0. Still, we found it helpful to use a separate instrument for gray, which partly reflects the fact that many languages have a separate name for it. The lightness l of gray or any other (combination of) colors is sonified as the pitch of the tone. Gray scale images, therefore will be sonified as a flute playing at varying pitch. Based on a musical scale, as shown in Fig. 11 (c), black, as the lowest lightness value, is assigned to the tonic keynote, whereas white to its octave. In between there are six whole tones and 11 semitones. For harmonic reasons we only utilize the whole tones of the octave and map each lightness value l between 0 and 1 to one of the eight tones. Further, we add thirds to all six intermediate tones. This creates a more comforting and aesthetical resonance and offers an elegant way to recognize whether one has reached the top or bottom of the scale, as they are played without thirds. Otherwise, users would need perfect pitch to recognize black and white. When working with scales in MIDI, each note has to be triggered and released, which, again, is why a very short *Attack-* and *Decay-*and as well a short *Release-Phase* is essential to maintain a close-to-continuous signal. In contrast, mixing colors on a constant luminance takes place solely within the *Sustain phase* for arbitrary time - the note itself does not change.

6.1.3. Calculation of Volume Shapes

Calculating the volumes of instruments in a mixture of sounds for all intermediate colors is an interpolation problem. Simple linear (barycentric) interpolation would be too restricted because once the overall volume of each instrument is set, there would be no way to counteract the dominance of some instruments in some specific mixtures. Therefore, we use *thin plate spline interpolation* [34] based on a set of control points. The fundamental idea behind the method is the physical model of a flat thin medal plate as in Fig. 12 (a) that is deformed by a few punctual strains, which we will call control values c. The plate is than forced into a new form that minimizes the

Figure 13: (a) Test set. (b) Possible orientations of objects.

deformation energy (Fig. 12 (b). We calculate a *Volume Shape* $\vartheta(h,s)$ that maps a volume ϑ to each color (h,s). For the bagpipe (color: yellow), this is visualized in Fig. 12 (c). The volume should be 100% at hue $h = 60°$ and full saturation $s = 1$, and 0% at hues h equal to 0° and 120° or greater, disregarding any saturation s. To control the volumes in mixed sounds, we add control values c in new positions (h_c,s_c). The calculation of such volume shapes $\vartheta(h,s)$ involves linear combinations of *radial basis functions* $f(h,s)$ [34]:

$$\vartheta(h,s) \approx \sum_{i=1}^{N} \lambda_i f_i(h,s) \quad (4)$$

where λ_i represents weighting of each $f_i(h,s)$ involved. These weights are found by minimizing a cost function that involves the sum of squared distances to all control values c_i, as well as the integral of the squares of the second partial derivatives of $\vartheta(h,s)$, serving as a smoothness term:

$$E = \sum_{i=1}^{N} \|c_i - \vartheta(h,s)\|^2 + \iint \left[\left(\frac{\partial^2 \vartheta}{\partial h^2} \right)^2 + \left(\frac{\partial^2 \vartheta}{\partial h s} \right)^2 + \left(\frac{\partial^2 \vartheta}{\partial s^2} \right)^2 \right] dh ds \quad (5)$$

6.2. Sonification of Texture Information

Sonifying both texture and color information is challenging in many ways. On the one hand we have to make sure that simultaneously played information is distinguishable, on the other hand we want to maintain a pleasing sound.

6.2.1. Acoustical Reliefs - Hearing Orientation Maps

We decided to utilize four more instruments, playing an octave below our keynote, to represent orientation maps of $\theta = \{0°, 45°, 90°, 135°\}$. The instruments we chose: Didgeridoo (0°), wood percussion (45°), uilleann pipes (90°), metal percussion (135°), create a hum alike sound at 0° and 90° and a percussion sound at 45° and 135°, to quickly distinguish horizontal and vertical from diagonal structures. Again, the framework allows exchanging instruments according to personal taste. To avoid *auditory masking* [35] we should guarantee that the Volume V_θ of all orientation map instruments are always lower than $\vartheta(h,s)$ for all h *and* s. The four orientation maps in between $\theta = \{22.5°, 67.5°, 112.5°, 157.5°\}$ are expressed using combinations of two neighbored *Orientation map* instruments, both playing at 50 % V_θ.

$V_{0°}(x,y) = d_4(x,y)$, $V_{22.5°}(x,y) = d_5(x,y)$, $V_{45°}(x,y) = d_6(x,y)$,
$V_{67.5°}(x,y) = d_7(x,y)$, $V_{90°}(x,y) = d_8(x,y)$, $V_{112.5°}(x,y) = d_9(x,y)$,
$V_{135°}(x,y) = d_{10}(x,y)$, $V_{157.5°}(x,y) = d_{11}(x,y)$

6.2.2. Audible Roughness - Sonification of Micro-Textures

Micro-Textures are sonified using one more instrument that has a vibrant temper. As [36] pointed out "a good *vibrato* is a pulsation of pitch, usually accompanied with synchronous pulsations of loudness and timbre, of such extent and rate as to give a pleasing flexibility, tenderness, and richness to the tone", which is an intuitive way to represent roughness acoustically.

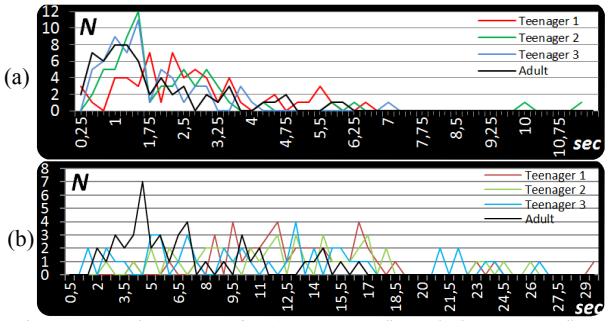


Figure 14: Histograms for (a) Exp. 1a/b and (b) Exp. 2a/b. N elements (y – *axis*) recognized in how many sec. (x – *axis*) each.

As anisotropic rough structures are visually salient and rarely occur in most environments, they are sonified more vibrant and at a Volume V_{Micro} being louder:

$$V_{Micro}(x,y) = \begin{cases} 0\%, & \text{if } d_{12}(x,y) = 0 \quad (\textit{smooth surface}) \\ 50\%, & \text{if } d_{12}(x,y) = 2 \quad (\textit{isotropic roughness}) \\ 100\%, & \text{if } d_{12}(x,y) = 1 \quad (\textit{anisotropic roughness}) \end{cases}$$

7. USER STUDIES

We did user studies on two groups of participants, following different motivations. First, as a proof of concept of our framework, we asked a congenital blind, 54 year old adult academic, who had acquired a geometric understanding and spatial sense throughout his life to solve several tests after 4 hours of training with our system. The participant was to solve three naming tasks at increasing difficulty:

- *Experiment 1a* was about identifying one out of four elements (orange, tomato, apple and lemon – as in Fig. 13 (a)) only by color while sonification of orientation maps and micro-textures was deactivated. Note that the target objects used for the task have the same spherical shape. In each of 60 trials, one of the 4 objects was selected at random and displayed at an arbitrary position on the touch screen. This was achieved by selecting one out of 40 images (10 per object, with the object in different positions) at random. The task of the participant was to find and name the object. In the evaluation, we focus on the time between the moment when the participant finds the object (which depends on where he starts and is therefore not very informative), and the moment when he names the object verbally to the experimenter (Table 1 and Fig. 14). The average time to simply find an object's position on the screen was about 1.7 seconds. Chance level (pure guessing) is 25% in this experiment.
- *Experiment 2a* involved orientation maps and color. This time, the participant had to recognize one out 7 objects (orange, tomato, apple, banana, cucumber, carrot, lemon), as shown in Fig. 13 (a), so both color and shape are important for correctly naming the object. Again, each element was presented individually (chance level: 14%) at arbitrary positions and also in one of eight orientations, as illustrated in Fig. 13 (b). The database consisted of 56 images (8 for each element, varying position and orientation). Again, times were measured between finding and naming the object verbally, as shown in Fig 14 and Table 1.
- *Experiment 3* was about recognizing an object within a set of other objects. Therefore, we presented images like the

one shown in Fig. 13 (a) on the touch-screen. In our database of 7 images, we made sure that two objects of equal color (e.g. banana and lemon) would not be positioned next to each other. In each trial, an image was presented and the participant was told, which object he had to find, based on a random generator. This time, we measured the overall time until an element was named.

In Experiment 1b and 2b, we tested the system on a group of congenital blind 14 year old teenagers. Unlike the adult participant, they had little geometric understanding and sense of space. We hope that our system can not only support them in everyday life, but also help them to develop cognitive abilities in geometry and spatial orientation. We performed an experimental evaluation of our system to measure their progress and compare it with the results of the adult participant. Three teenagers were trained about 5 hours with the system. This also included fundamental lecturing about basic geometry. Then, we asked them to perform Experiment 1b and 2b, which had the same setup as 1a and 1b described above. Surprisingly, the teenagers were able to perform the tests with similar hit rates and times as our adult participant (Table 1 and Fig. 14).

RESULTS	**P**	**Hit rate**		$\tilde{X}$	μ	σ
Experiments	**-**	**%**	**elem.**	**sec**	**sec**	**sec**
Exp. 1a	**A**	**100.0**	**60/60**	**1.3**	**1.8**	**1.4**
	T_1	91.6	55/60	2.2	2.6	1.5
Exp. 1b	T_2	93.3	56/60	1.5	2.3	1.9
	T_3	100.0	60/60	1.3	1.7	1.1
Exp. 2a	**A**	**93.3**	**42/45**	**5.6**	**7.0**	**3.9**
	T_1	88.8	40/45	12.1	13.3	4.7
Exp. 2b	T_2	93.3	42/45	11.9	12.5	5.5
	T_3	88.8	40/45	10.1	11.4	6.6
Exp. 3	**A**	**100.0**	**45/45**	**5.6**	**10.6**	**12.0**

Table 1. Hit rates and times (median $\tilde{X}$, mean μ, and standard deviation σ), for each trial and participant *P*.

8. CONCLUSIONS

We have presented a general framework and a sample implementation of a device that can support blind and visually impaired persons in exploring images or scenes. Many details of our implementation, such as the choice of local image descriptors, may be modified or improved further. However, we tried our best to design descriptors that are most promising from the theoretical and most informative from the practical point of view. The same is true for our sonification concepts: they are only one way how this can be achieved, yet we argue that it is an appropriate and powerful way to do it. The experimental results indicate that the system enables users to solve simple recognition tasks fast and reliably. In future experiments, we are planning to consider more and more difficult tasks with cluttered scenes and a wider variety of objects. Both the design of image descriptors and sonification concepts went through many experimental steps, and we discarded many alternative designs before we ended up with the solution that we present here. Feedback from the users was that they found the setup that we presented in this paper both intuitive and helpful. Still, it is our goal to start a fruitful discussion about 1: which features of an image are most informative in this framework, and 2: how can sonification convey as much relevant information to the user as possible. As we mentioned in the paper, there are also many possible extensions in terms of sensorics (cameras, tracking systems) and exploration paradigms. In future work, we are planning to continue to improve and extend our system along these lines. Our vision is to provide visually impaired persons with software for web browsers, image "viewers", and on portable systems such as smart phones.

9. ACKNOWLEDGEMENTS

We thank the Rheinischen Blindenfürsorgeverein, Düren, especially Marina, Larissa, Florian, Sascha and Mrs. Gut for their interest and participation in the project. We also thank Tobi and Rainer for their highly appreciated advisory support.

10. REFERENCES

[1] G. Kramer, B. Walker et. al.,"Sonification Report." in *ICAD,* 1999.
[2] F.Grond, S.Janssen et.al.,"Browsing RNA structures by interactive sonification," in *Proc. of ISon, 3rd ISW,* Stockholm, 2010.
[3] K. Vogt et al., "A Metaphoric Sonification Method – Towards the Acoustic Std. Model of Part. Physics," in *16th ICAD.,*USA,2010.
[4] J. Xu et al., "Sonification based Electronic Aid System for the Visually Impaired," in *JCIT,* vol. 6, no. 5, 2011.
[5] P. Meijer, "An experimental System for Auditory Image Representation," in *IEEE TBME.,* vol. 39, no.2, pp. 112-121, 1992.
[6] D. Margounakis et. al., "Converting Images to Music using their Color Properties," in *12th Int. Conf. on A.D., London* , 2006.
[7] K. Van den Doel, "Sound View: Sensing Color Images by Kinesthetic Audio," in *9th Int. Conf. on A.D.,* Canada , 2003.
[8] D. Payling et al., "Hue Music"in *13th Int. Conf. on A.D.,*CA, 2007.
[9] T. Hermann, "Taxonomy and Definitions for Sonification and Auditory Displays," in *Pro. 14th Int. Conf. on A.D.,* France, 2008.
[10] A. Hunt, T. Hermann, "The Importance of Interaction in Sonification," in *Pro. 10th Int. Conf. on A.D.,* Australia, 2004.
[11] D. Wagner et al., "ARToolKitPlus for Pose Tracking on Mobile Devices," in *Proc. of 12thCVWW,* Austria, 2007.
[12] E. Goldstein, *Sensation and Perception*, C. L. Emea, 2009.
[13] M. Lew, *Princ. of Visual Information Retrieval*, Springer, 2001.
[14] C. Tomasi, R. Manduchi, "Bilateral Filtering for Gray and Color Images," in *Proc. of Int. Conf. on CV,* India, 1998.
[15] J. Canny, "A Computational Approach to Edge Detection," in *IEEE TPAMI,* vol. 8, no. 6, pp. 679-698, 1986.
[16] B. Mandelbrot, *The Fractal Geometry of Nature*, H. Holt, 2000.
[17] B. Jähne, *Digital Image Processing,* Berlin: Springer, 2005.
[18] E. H. Adelson, P. J. Burt, "The Laplacian Pyramid as a Compact Image Code," in *IEEE TC,* pp. 284-299, 1983.
[19] D. Gabor., "Theory of comm.,"in *J. IEEE,*vol.93,pp.429-459,1946.
[20] M. Zhou, H. Wei, "Face Verification using Gabor Wavelets and Ada Boost," in *Int. Conf. on P.R.,* pp. 404-407, 2006.
[21] S. Suzuki et al.,"Top. structural analysis of digitized binary images by border following," in *CVGIP,* vol. 32, no. 1, pp. 32-46, 1985.
[22] R. Sedgewick, *Algorithms,* Addison Wesley, 1992.
[23] R. Haralick, "Statistical and structural approaches to texture," in *Proc. IEEE,* vol.67,no. 5, pp. 786 – 804, 1979
[24] M. Gippet, et al., "Haralick's texture features computed by GPU's for biological applications," in *IJCS*, 2009.
[25] P. Nelson, *Biological Physics*, Palgrave MacMillan, 2007.
[26] Y. Boykow, et al., "Efficient Approximate Energy Minimization via GraphCuts,"in *IEEE TPAMI,*vol.20,no.12,pp.1222-1239, 2001.
[27] Y. Kolmogorov, et al., "What Energy Functions can be minimized via Graph Cuts," in *IEEE TPAMI,*vol.26,no.2,pp.147-159, 2004.
[28] Y. Boykow, Y. Kolmogorow, "An Experimental Comparison of Min-Cut/Max-Flow Algorithms for Energy Minimization in Vision," in *IEEE TPAMI,* vol. 26, no. 9, pp. 1124-1137, 2004.
[29] A. Rosenfeld, *Dig. Picture Processing*, Acad.Press, Orlando, 1982.
[30] S. Yeo, J. Berger, "Raster Scanning", in *Proc of ICMC,* 2006.
[31] T. Yoshida et. al., "EdgeSonic", in *Augm. Human Int. Conf.,* 2011.
[32] R. Ramlollet. al., "Constructing sonified haptiv line graphs for the blind student: first steps," in *4th Int. ACM CAT,* 2000.
[33] M. Russ, *Sound Synthesis and Sampling,* Focal Press, 2008.
[34] G. Donato, S. Belongie, "Approximation Methods for Thin Plate Spline Mappings and Principal Warps,"in *Proc of LNCS*,DK,2003.
[35] P. Gray, *Psychology,* Worth Publishers Inc., 2002.
[36] C. Seashore, Studies in the psychology of music Vol. 1: The vibrato, University of Iowa City, 1938.

SONIFYING ECOG SEIZURE DATA WITH OVERTONE MAPPING: A STRATEGY FOR CREATING AUDITORY GESTALT FROM CORRELATED MULTICHANNEL DATA

Hiroko Terasawa[†]*, Josef Parvizi*[‡]*, and Chris Chafe*[‡]

[†]University of Tsukuba / JST-PRESTO
1-1-1 Tenno-dai, Tsukuba
Ibaraki 305-8577 Japan
terasawa@tara.tsukuba.ac.jp

[‡] Stanford University
Stanford, CA 94305 USA
jparvizi@stanford.edu
cc@ccrma.stanford.edu

ABSTRACT

This paper introduces a mapping method, *overtone mapping*, that projects multichannel time-series data onto a harmonic-series structure. Because of the common-fate effect of the Gestalt principle, correlated signals are perceived as a unity, while uncorrelated signals are perceived as segregated. This method is first examined with sonification of simple, generic data sets. Then overtone mapping is applied to sonification of the ECoG data of an epileptic seizure episode. The relationship between the gestalt formation and the correlation in the data across channels is discussed in detail using a reduced 16 channel data set. Finally, sonification of a 56-channel ECoG data set is provided to demonstrate the advantage of the overtone mapping.

1. INTRODUCTION

We report a method to represent correlation structures of multichannel signals. This method, called *overtone mapping*, projects large-scale, multichannel data onto a harmonic series (a.k.a. an overtone series) of a sound, and the correlated elements across channels are perceived as a fused "auditory gestalt[1]." The benefit of the method is that humans can intuitively perceive the similarity patterns across channels in the data without statistical analyses. We first describe the principle of this method, and then we introduce the application for electrocorticography (ECoG) data during an epileptic seizure episode.

A harmonic series is a commonly found structure in voices and instrumental sounds. We perceive harmonic series as a single, integrated stream of sound, when they share common fate, and their temporal deviations are perceived as deviations in timbre. Using this property, we could present a set of independently measured data channels as a coherent auditory unit when the data share a common fate across channels–i.e., when the data are correlated.

Although this work might seem to be just another example of parametric mapping sonification of brain-wave data, in addition to the previously introduced, sophisticated sonification examples [1, 2, 3], we trust that this work contributes to finding a design-by-principle method for perceptually meaningful sonification. Readers might recall the problems Flowers pointed out in his paper in ICAD2005 [4] as "things we need to know more about." In this section, he questions the role of timbre in stream segregation, and seeks a method to monitor two or more processes that are co-occurring in real time. We believe this paper answers some of these questions. This is also an example of using timbre as a medium for projecting the complexity of data, as urged in the notable publications [5, 6, 7].

In this paper, we explore the gestalt principle and the effect of overtone mapping by theoretical considerations and sound examples, rather than merely conducting a user-evaluation test. We request that our readers spend some time exploring the sound examples provided online. All of the sound examples that we discuss in this paper are uploaded on this Website:
http://www.tara.tsukuba.ac.jp/%7Eterasawa/ICAD2012/

In the following sections, we first briefly review the common-fate principle in gestalt perception. Then we describe overtone mapping by generic data examples and apply overtone mapping in ECoG data sonification.

2. AUDITORY GESTALT FORMATION AND THE SONIFICATION OF CORRELATED DATA

2.1. Auditory gestalt and its principles

Gestalt perception is the perception of a specific whole or unity, by integrating its parts. Similar to the visual domain, gestalt perception also occurs in the auditory domain. The phenomenon of auditory gestalt is well discussed in "Auditory Scene Analysis" by Bregman [8]. The formation of gestalt perception is described by several principles. Elements such as proximity, symmetry, similarity, continuation, closure, and common fate contribute to the perceptual organization.

2.2. Common fate shared across harmonic series produces a perception of unity

Among those, the common-fate effect was well-investigated in the writing and compositions by John Chowning. He introduced how to form auditory gestalt in terms of the common-fate principle [9, 10, 11]. On a harmonic series of sinusoids, he applied subtle frequency modulations (micro-modulation) at a few different modulation frequencies that mimick vibrato, with some overtones at one vibrato frequency and some other overtones at another vibrato frequency. As a result, the sinusoids that were modulated with the same vibrato frequency became perceived as a unity, and a few voices can exist simultaneously in a stream. In other words,

[1] In this paper, we use "auditory gestalt" meaning "a perceived auditory unity," and "Gestalt principle" meaning the grouping law proposed by German Gestalt school of psychology.

the "common fate" in Chowning's examples is afforded by sharing the same vibrato frequency among harmonic series. Using this technique, he was able to render gradually arising vibrato voices out of a static sinusoidal superposition. This effect is well employed in his pieces *Phoné* (1980-1981), and *Voices* (2005).

2.3. The correlation across channels can function as common fate for an auditory gestalt

Formation of a unity perception by the harmonics sharing common fate provides a good opportunity for data sonification of multichannel, correlated, time-series data. In multichannel time-series data, such as electromyograph (EMG), electroencephalogram (EEG), and electrocorticography (ECoG), the acquired data are often strongly correlated across channels. The similarity analysis, or any other kind of statistical analysis of the correlated yet separately measured time-series data is computationally demanding. Using the common-fate effect, in other words, interpreting the correlation as a common fate, we can easily present the correlated data as a perceived unity, arising out of uncorrelated elements, without applying statistical analysis beforehand.

3. OVERTONE MAPPING WITH GENERIC DATA

In this section, we describe the formation of auditory gestalt by the common-fate effect using generic data and their sonification examples. The sound examples are provided as sounds 1-6 on the Website. Readers are strongly recommended to listen to these sounds to experience the auditory gestalt formation by the common-fate effect.

3.1. Sound 1: Harmonic series with sinusoidal amplitude modulation

This is the reference pattern for the rest of the examples. Figure **??** shows the amplitude pattern for the time course of this sound. The fundamental frequency is 440 Hz, and the sound has eight harmonics (i.e., overtones at integer-multiples of the fundamental frequency). Each of eight harmonics is amplitude modulated with a sinusoidal pattern of a single modulation frequency. Sharing a single modulation pattern, all the harmonics are perceived as unity.

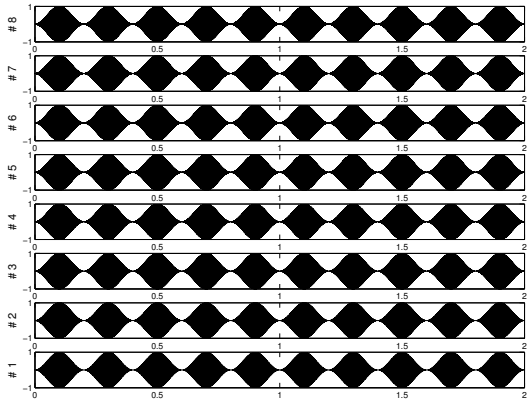


Figure 1: Sound 1. Each row in the figure shows the amplitude pattern over time of each harmonic, from the 1st to the 8th harmonics from the bottom to the top row, respectively. This example has the same sinusoidal amplitude pattern for all the eight harmonics.

3.2. Sound 2: Static and sinusoidal patterns

In Sound 2, the modulations of the 3rd, 6th, and 7th harmonics are removed as shown in Fig. 2. Now these harmonics with a static pattern are perceptually segregated, forming another unity of static tone. The rest of the harmonics with sinusoidal modulation forms another unity. The degree of segregation is moderate compared with some of the following examples.

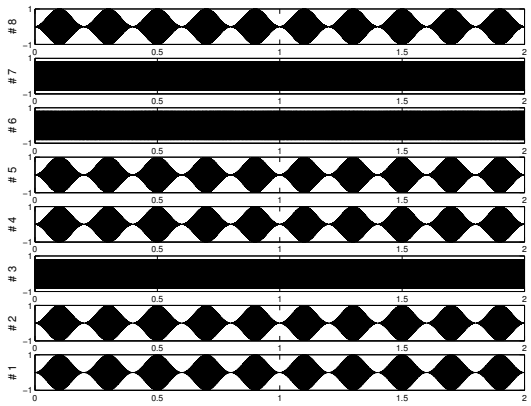


Figure 2: Sound 2. The 3rd, 6th, and 7th harmonics are static without modulation, providing a static tone unity.

3.3. Sound 3: Sinusoidal patterns with two frequencies

In Sound 3, the modulations of the 3rd, 6th, and 7th harmonics are slower, as shown in Fig. 3. These harmonics with the slower modulation pattern are perceived segregated forming a clear unity. The rest of the harmonics with sinusoidal modulation form another unity.

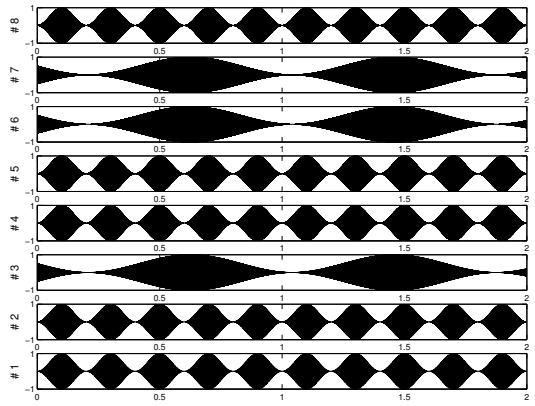


Figure 3: Sound 3. The 3rd, 6th, and 7th harmonics are modulated with a slower modulation frequency.

3.4. Sound 4: Chirp-like and sinusoidal patterns

Sound 4 provides a dynamic transition in the temporal pattern as shown in Fig. 4. The frequency of amplitude modulation at the 3rd, 6th, and 7th harmonics increases over time, forming a chirp-like pattern. When two modulation frequencies (one for 3, 6, 7, and another for the rest) are very distant, the segregation is easier. However, when the two modulation frequencies are crossing, all the harmonics are perceived fusing into a unity.

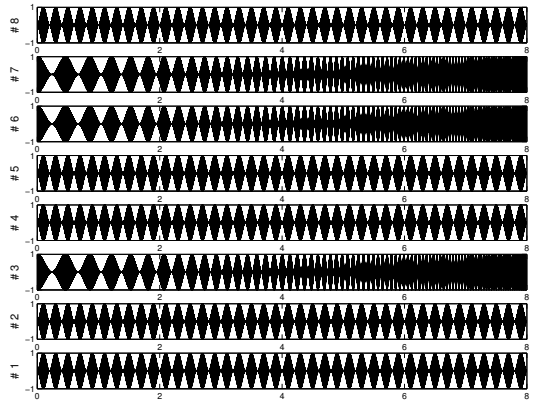


Figure 4: Sound 4: The modulation frequency for the 3rd, 6th, and 7th harmonics increases over time.

3.5. Sound 5: Non-sinusoidal and sinusoidal patterns

So far, we have considered only sinusoidal and static patterns. This example, Sound 5, provides the case that a temporal pattern does not need to be sinusoidal. As shown in Fig. 5, the 3rd, 6th, and 7th harmonics now share a pattern of decaying amplitude. When we hear this sound, these harmonics are perceived as a quickly decaying unity, against the sinusoidally modulated unity of the rest. This segregation is clearly perceived.

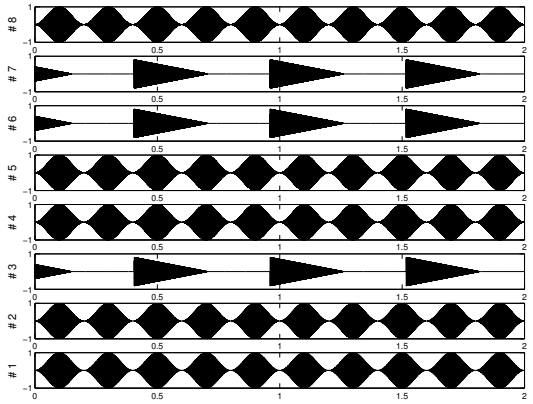


Figure 5: Sound 5: The 3rd, 6th, and 7th harmonics share the decaying-amplitude pattern.

3.6. Sound 6: Sinusoidal patterns with a phase difference

After considering the patterns varying with their duration, it is now worthwhile seeing whether we could create segregation just by changing the phase of the same sinusoidal pattern. Sound 6 provides such an example: the 3rd, 6th, and 7th harmonics are now presented with a $\pi/4$ phase difference from the rest of the harmonics, as shown in Fig. 6. The segregation is ambiguous yet noticeable. As the phase difference reaches the opposite (a difference of π), the segregation becomes slightly clearer. However the unities that differ only by their phase are easily confused.

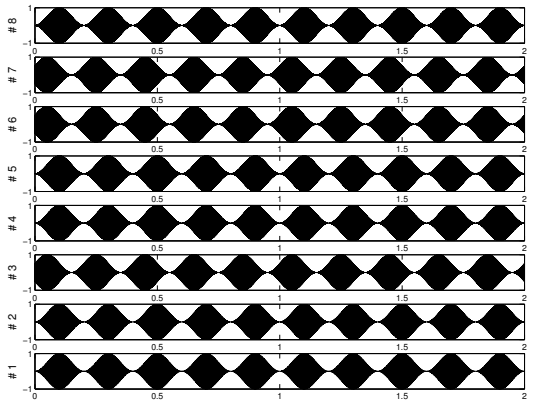


Figure 6: Sound 6: The 3rd, 6th, and 7th harmonics differ only by their phase from the rest of the harmonics.

3.7. Discussion of the generic data examples

In this section, we demonstrated the principle of auditory gestalt formation by common fate with the sonification of simple, generic temporal patterns. The more the temporal patterns differ from each other, the clearer the perceptual segregation is. These temporal patterns could be sinusoidal as having shown by Chowning's examples, or they could be non-sinusoidal patterns as provided in Sound 4 and 5 examples, as long as a set of harmonics shares the same common fate. The grouping by phase is noticeable but not prominent.

4. OVERTONE MAPPING APPLIED TO ECOG DATA

4.1. About the ECoG data

In this section, we consider the overtone mapping method applied to a set of ECoG signals. The ECoG measurement was done as a part of clinical procedure by Josef Parvizi at Stanford University Hospital, under the guidance of Stanford Institutional Review Board. The patient was personally consulted about the project and gave full consent. The original signals were measured with 56 channels, and the measurement lasted for many days. In this discussion, we focus on the excerpt of only 10 s. This excerpt captures a very interesting moment in the epileptic seizure episode, in which multiple channels show the mixture of coherent and non-coherent neural activities.

This excerpt for 56 channels is plotted in Fig. 7. These 56 channels show complex correlation patterns, to which we will return at the end of this section. However, in order to address the relationship between the correlation and common fate effect, 56 channels are just too many. Therefore, we decided to select some prominent channels out of 56. Figure 8 shows a stem plot of the mean absolute amplitude of the 56-channel data. As you can see from the figure, some of the signals are stronger than others, and we selected the 16 strongest mean-absolute-amplitude channels, assuming those strong channels carry more meaningful information with less measurement noise.

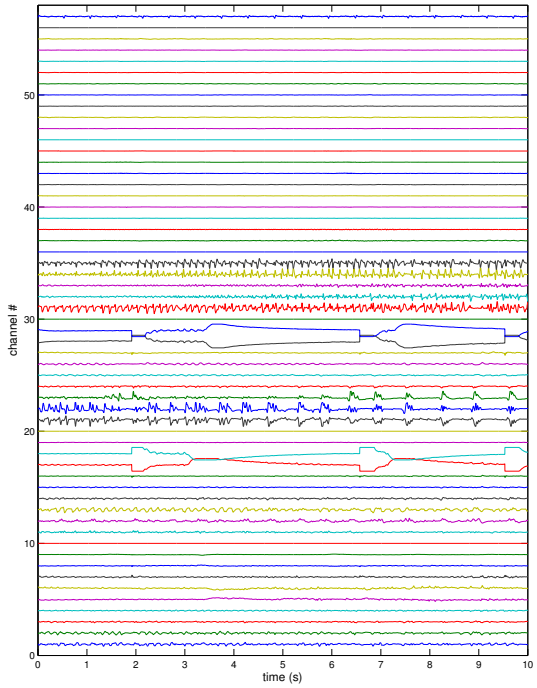


Figure 7: Plot of 56-channel ECoG data for 10 s. Each line shows the signal for each channel, from the bottom to the top showing channels 1 to 56, respectively.

4.2. Sonification of ECoG data

The sonification of the 16-channel excerpt data was done using the following procedure.

1. The fundamental frequency was set to 180 Hz.
2. Harmonics of 16 sinusoids (up to the 16th harmonics) were created.
3. Each harmonic was amplitude-modulated by each channel: the 1st harmonic is modulated with channel 1, the 2nd with channel 2, and so on.
4. All of the harmonics were summed, creating a single audio signal.
5. The audio signal was linearly scaled with its maximum value, so that the scaled signal could fit within the .wav file dynamic range.

The 16-channel ECoG sonification is available as "ECoG Sound 1" on the Website.

Listening to the sonified sound, we notice some clear patterns existing within the dynamically transitioning harmonic series, although the mapping was decided blindly without signal analysis.

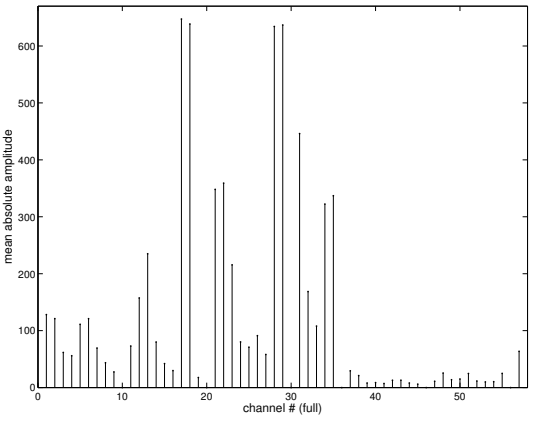


Figure 8: Mean absolute amplitudes of channels 1-56. 16 channels with strong amplitudes were selected for the following discussion.

Table 1: Groups of Correlated Signals

Group	Channels
1	1, 2, 3, 4, 5, 8, 9, 10
2	6, 7, 11, 12
3	13, 14, 15, 16

4.3. Discussion on the ECoG data sonification

When we listen to the 16-channel ECoG data sonification, we notice that there are a few recognizable gestalts, which can be identified with correlation analysis. Figure 9 shows the correlation matrix of 16 channel signals on 16 $\times$16 square color tiles. Each square at (n, m) position represents the value of correlation between the signals at channel n and channel m. By viewing this figure, we could find a few islands of more correlation–namely groups 1, 2 and 3–of the channels listed in Table **??**.

By creating subset-tones of the sonification, we can verify the formation of auditory gestalt. This could be done by replacing the step 4 of the procedure introduced in Section 4.2. Instead of summing all of the harmonics, we now sum only the harmonics that correspond to each group. Figure 10 shows the wave plot of each subset-tone for groups 1, 2, and 3. These subset-tones can be heard as ECoG sound 2-4 on the Website.

As verified in the waveform plot and sound examples, each group of correlated signals clearly forms an auditory gestalt, which is easily recognized. The recognizable patterns in the 16-channel sonification were the auditory unities arising from the correlated signal patterns.

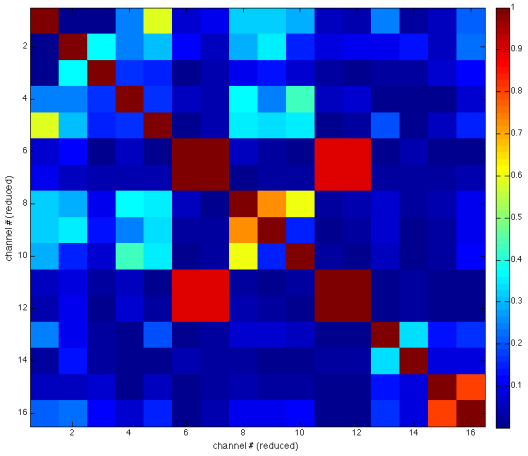


Figure 9: Correlation matrix of the selected 16-channel signals.

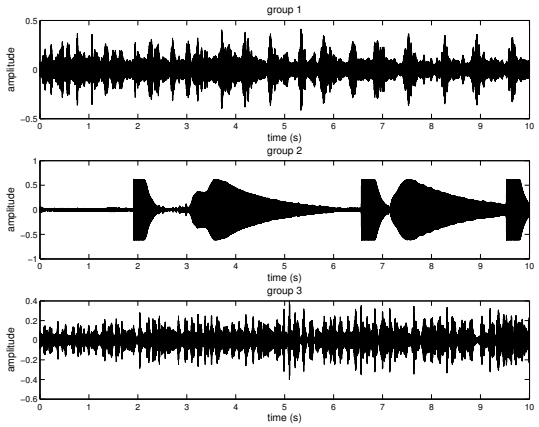


Figure 10: Waveform plot of subset-tones of sonification: Group 1 (top), group 2 (middle), and group 3 (bottom).

way as Fig. 9, but its correlation patterns are not easily recognizable. However, when we listen to the sonification of the 56 channels (fundamental frequency: 120 Hz; number of harmonics: 56), we can hear a handful of patterns with rich textures arising from the broad spectral components, in the same way as its 16-channel version. The 56-channel sound is provided as "ECoG Sound 5" on the Website.

The visual representation of the correlation is not trivial, but the auditory representation of the correlation by common fate effect is more recognizable.

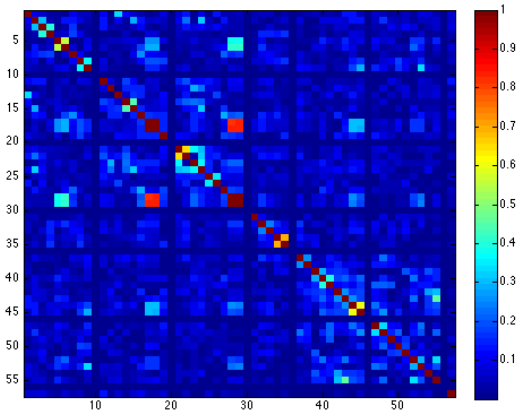


Figure 11: Correlation matrix of the full 56-channel signals.

4.4. Demo: 56-channel ECoG data sonification

Finally, we want to introduce the full-data example. However, analyzing the similarity in 56-channel signals becomes increasingly challenging. Figure 11 shows the correlation matrix in the same

5. CONCLUSION AND FUTURE WORK

In this paper, we discussed the formation of auditory gestalt by the common-fate principle. With the generic data sonification, we demonstrated that two distinct temporal patterns can be mapped to

amplitudes of harmonic series, and that this mapping can provide a auditory segregation. The degree of segregation–i.e. how clearly the auditory gestalts can be perceptually segregated–depends on the degree of similarity between the two temporal patterns. The temporal pattern could take any shape other than sinusoids, as long as it holds a distinct temporal pattern. In the later section of the paper, we introduced another example that applied the same mapping to real 56-channel ECoG data. With the reduced 16-channel version, we could see the clear correspondence between the data correlation and auditory gestalt formation by overtone mapping. Furthermore, the 56-channel version serves as an example that auditory gestalt formation is much easier and simpler than statistical analysis of the data similarity across many channels. The advantage of overtone mapping is that our auditory perception can easily judge the similarity of the signals across channels.

In this paper, we presented the gestalt formation by overtone mapping by conceptual and theoretical considerations and by sound examples. Quantitative formalization of this technique remains as a future consideration. Overtone mapping seems to be a useful approach not only for ECoG signals but also for EEG and EMG signals. Investigating the applications for these, and other types of signals would be desirable in the future. Finally, while this paper describes the auditory gestalt formation using the common-fate principle, another paper by the first author on the sonification of the genetically modified *C. Elegans* [12] provides an example for the gestalt formation by proximity principle. Investigating the sonification according to the rest of the principles (i.e., symmetry, similarity, continuation, and closure) will enable further theorization of the auditory gestalt formation in data sonification.

6. ACKNOWLEDGMENT

The author would like to thank John Chowning for his inspiring works and presentations. The sonification of ECoG signals was conducted during Hiroko Terasawa's doctoral study at the Center for Computer Research in Music and Acoustics (CCRMA), Stanford University. The authors wish to thank the anonymous patient who contributed this unique data set for the epileptic research. This work is supported by JST-PRESTO program.

7. REFERENCES

[1] T. Hermann, P. Meinicke, H. Bekel, H. Ritter, H. M. Muller, and S. Weiss, "Sonifications for EEG data analysis," in *Proceedings of the 2002 International Conference on Auditory Display, Japan.*, 2002.

[2] G. Baier, T. Hermann, and U. Stepheni, "Event-based sonification of EEG rhythms in real time," *Clinical Neurophysiology*, vol. 118, pp. 1377–1386, 2007.

[3] T. Hermann, G. Baier, U. Stephani, and H. Ritter, "Kernel regression mapping for vocal EEG sonification," in *Proceedings of the 2008 International Conference on Auditory Display, France.*, 2008.

[4] J. H. Flowers, "Thirteen years of reflection on auditory graphing: Promises, pitfalls, and potential new directions," in *Proceedings of the 2005 International Conference on Auditory Display, Ireland.*, 2005.

[5] S. Barrass and G. Kramer, "Using sonification," *Multimedia Systems*, vol. 7, pp. 23–31, 1999.

[6] G. Kramer, B. Walker, T. Bonebright, P. Cook, J. Flowers, N. Miner, and J. Neuhoff, "The sonification report: Status of the field and research agenda," tech. rep., Prepared for the National Science Foundation by members of the International Community for Auditory Display Editorial Committee and Co-Authors, 1999.

[7] F. Grond and J. Berger, *The Sonification Handbook*, ch. 15. Parameter Mapping Sonification. Logos Publishing House, Berlin, Germany, 2011.

[8] A. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press, 1994.

[9] J. Chowning, "Synthesis of the singing voice by means of frequency modulation," in *Current Directions in Computer Music Research* (M. V. Mathews and J. R. Pierce, eds.), MIT Press, 1989.

[10] J. Chowning, "Music from machines: Perceptual fusion and auditory perspective – for ligeti," tech. rep., Stanford University Department of Music Technical Report STAN-M 64, 1990.

[11] J. Chowning, *Music, Cognition, and Computerized Sound: An Introduction to Psychoacoustics*, ch. Perceptual Fusion and Auditory Perspective. MIT Press, 1999.

[12] H. Terasawa, Y. Takahashi, K. Hirota, T. Hamano, T. Yamada, A. Fukamizu, and S. Makino, "C. elegans meets data sonification: Can we hear its elegant movement?," in *Proceedings of the 8th Sound and Music Computing Conference*, 2011.

RECOGNITION OF AUDIFIED DATA IN UNTRAINED LISTENERS

Robert L. Alexander, Sile O'Modhrain, Jason A. Gilbert, Thomas H. Zurbuchen

University of Michigan,
Ann Arbor, MI, 48104
robertalexandermusic@gmail.com

Mary Simoni

Rensselear Polytechnic University
School of Humanities Arts and Social Sciences
msimoni@rpi.edu

ABSTRACT

The effective navigation and analysis of large data sets is a persistent challenge within the scientific community. The objective of this experiment was to determine whether participants who received no training were able to identify audified data sets at a rate above chance in a forced-choice listening task. Nineteen participants with various levels of musical and scientific expertise were asked to place audio examples into one of the five following categories: Digitally Generated Sound - White Noise, Solar Wind Data, Neuron Firing Data from a Human Brain, Seismic Data (Earthquake Activity), and Digitally Generated Sound - Sinusoidal Waveform. At no time were participants made aware of the accuracy of their responses during the experiment. Participants who had never been exposed to audified data sets were able to recognize audification examples at a rate that was 23 percentage points above chance performance; however, the sample size of individuals with no previous exposure to audified data was not large enough to determine statistical significance. When controlling for previous exposure to any of the provided listening examples, all participants performed well above the statistical likelihood of chance responses in the recognition of digitally generated white noise and sinusoidal waveforms. This indicates that participants with no previous exposure to audified data were able to discriminate between audified data and digitally generated sounds.

1. INTRODUCTION

Sonification is the science that concerns the transfer of information through sound. The *Sonification Report* broadly defines this term as "the use of non-speech audio to convey information." [1] Audification is a specific form of auditory data analysis in which data samples are isomorphically mapped to audio samples. This method has proven successful in uncovering new insights that would otherwise be overlooked through traditional analysis methods [2, 3]. However, no methodological framework has been established for how this process may be successfully implemented for exploratory data analysis across a wide range of scientific disciplines. One goal of this experiment is to establish a baseline measurement for human ability to recognize audified data sets.

Formal research in the field of auditory data analysis can be traced back to the year 1946, when a volume was published on the *Principles of Underwater Sound* with the goal of advancing sonar techniques [4]. Three years later *The Mathematical Theory of Communication* laid the foundation for our modern understanding of signal processing techniques [5]. Early auditory display research demonstrating that multi-modal stimulation could greatly increase the rate of information transfer to a human operator [6-8]. This investigation was later extended to human pattern matching abilities, finding that known visual-analysis methods were often inferior to auditory analysis in the representation of multivariate data [9]. Several additional experiments utilizing multivariate data were conducted by Bly, and it was noted that "sound can indeed increase the information about multivariate data when it is presented to a human analyst." [10, 11]

Sonification techniques have been employed in a wide range of scientific studies that build upon these early foundations. In *An Illustrated Analysis of Sonification for Scientific Visualization* it was noted that, "all aspects of sonic display of information need further research." A discrete set of possible areas where sonification research could be beneficial were offered, including: data representation, interaction processes, and validation of graphical processes [12].

Modern auditory data analysis techniques are commonly taught in academic settings, though this instruction is geared towards expertise in music-production. A course at the University of entitled "Timbral Ear Training" teaches students to notice subtle changes in the spectral composition of white and pink noise fields [13]. It is possible that this type of training could also prove effective in enabling researchers to recognize subtle differences between audified data sets. The objective of this experiment is to determine a pre-training baseline rate for successful recognition of audified data sets, with a comparison against chance performance utilized as a metric. Audified data sets will be presented in conjunction with digitally manufactured noise and sinusoidal waveforms, as previous research has suggested that auditory data analysis can be beneficial in the identification of equipment-induced noise [2].

2. METHODS

2.1. Participants

Nineteen participants took part in this experiment (6 female, 13 male; age 21 to 40). A pre-test questionnaire established that four participants had received a high school diploma, ten had received a bachelor's degree, and five participants had received a Masters or PhD. Three participants

had no musical training, one participant had a single year, four participants had two to three years, seven participants had four to six years, and four participants reported seven or more years of musical training.

All but three participants self-reported average to above-average hearing. A single-frequency auditory threshold test was administered after the listening portion of the survey. This test provided 300ms bursts of a 440hz sinusoidal waveform spaced evenly at 1 second intervals. The gain of each successive waveform was reduced by 6db. Individuals who self-reported below average hearing showed no statistically significant difference in performance on this task ($P < 0.22$).

A post-test questionnaire revealed that of the nineteen participants, thirteen had previously been exposed to audified data in one form or another. All participants reported average or above average expertise with computers, and ten reported average or above average computer-music expertise (nine reported below average). All but two participants reported experience with data analysis, mathematical modeling, and/or scientific research.

2.2. Procedure

The experiment was administered within a custom software-interface built in the Max/MSP programming environment (see **Figure 1**). After completing a short pre-test questionnaire, participants were asked to listen to a series of audio examples played back over headphones. Participants were verbally informed that they could either push a button with the mouse, or press the space bar to play back audio examples. Before beginning the listening task, participants were provided with a spoken-word listening example, and asked to set their audio-playback to a comfortable level utilizing a volume-slider provided within the software interface. The participants' task was to correctly identify the source of a sound from a list of five choices. This forced-choice task included the following available responses for all listening examples: Digitally Generated Sound - White Noise, Solar Wind Data, Neuron Firing Data from a Human Brain, Seismic Data (Earthquake Activity), and Digitally Generated Sound - Sinusoidal Waveform.

On-screen instructions informed participants that audio files were either generated from scientific data or digitally manufactured. It was also made clear that multiple examples of each type could appear over the course of the experiment. A total of 8 audio files were utilized for the listening task, these included two examples of audified neuron firing data from a human brain, two examples derived from solar-wind data, two examples of audified earthquake data, and one example of both white noise and a sinusoidal waveform. Each audio example was provided 3 times: Once at full speed, once at 75% full speed, and once at half speed. Twenty-four listening examples were provided in total.

Participants were asked to make their best guess as to the source of the audio, and then press a separate button labeled "submit" to enter their selection. At no time were participants made aware of the accuracy of their selection during the experiment (the experimental coordinator was always present within the room, but did not answer any questions pertaining to accuracy of responses). Participants were provided with a number corresponding to the current question (out of 46 total questions), such that they could track their progress towards completion. An on-screen clock began counting upwards at the beginning of the pre-test questionnaire. An on-screen level-meter provided visual feedback as to the volume of the audio file at 50ms intervals.

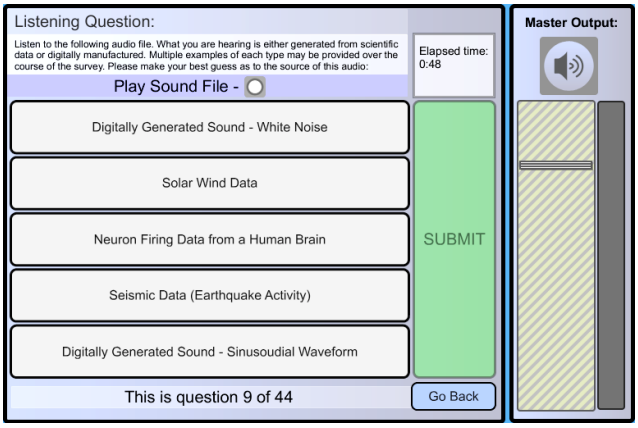


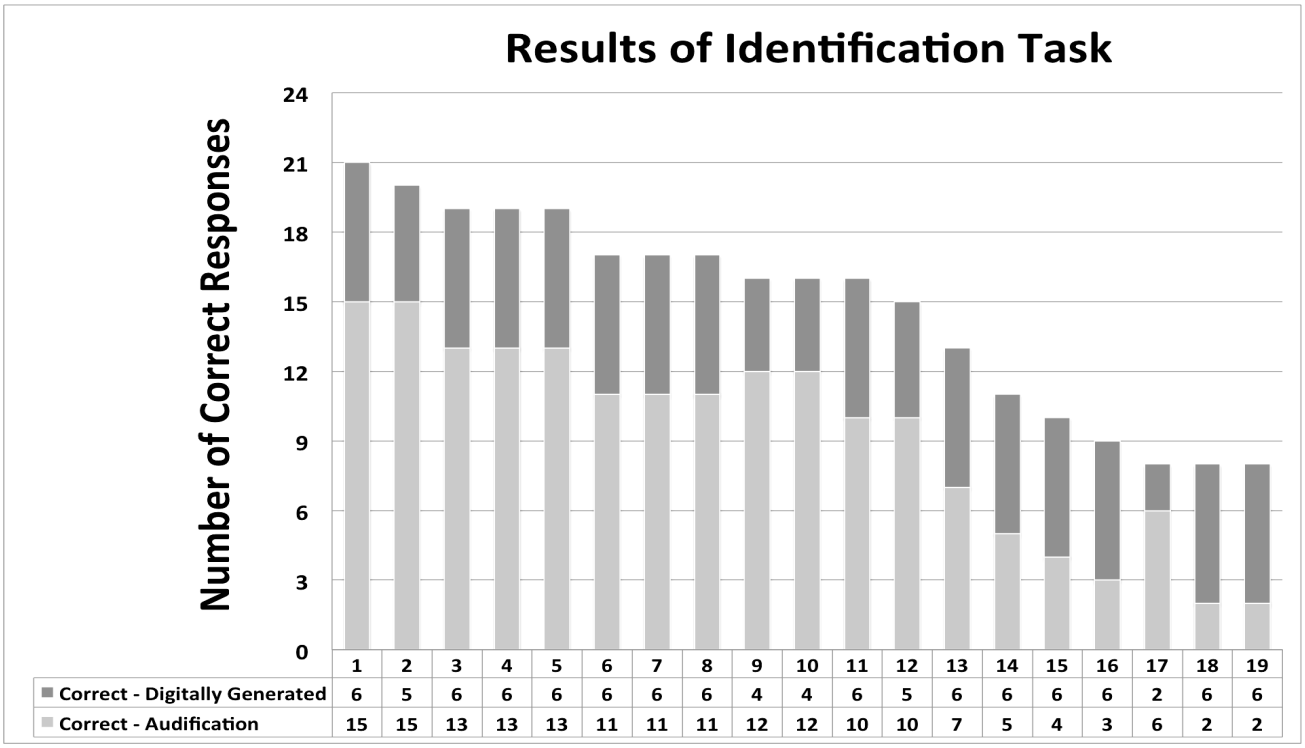
Figure 1. Participants were provided with a list of potential audio sources and asked to guess which example they were currently listening to.

2.3. Stimuli

A total of 8 audio files were utilized for the listening task, all files were 16-bit AIFF format at a sampling rate of 44.1kHz. The seismic data files were downloaded from a server in an audified data format (.wav). The solar wind and neuronal-firing examples were audified with a novel algorithm in the Matlab programming environment. This algorithm transferred the original comma-separated data files into 2-dimensional arrays, and then determined minimum and maximum values in each column of data. These values were utilized to scale the data as floating-point values between -1 and 1. These values were then sequentially mapped to 16-bit audio samples with the "wavwrite" function (all files were ultimately converted to AIFF format for playback in Max/MSP).

All audio files in this experiment were balanced to a similar playback amplitude (RMS). Examples ranged from approximately one to eight seconds in length, with a mean length of 5.3 seconds. All samples (except for the seismic data) were smoothly faded in and out over the course of approximately one to two seconds. A total of eight audio files were utilized for the listening task, each audio example was played back at total of three times: once at full speed, once at seventy-five percent of the full playback speed, and once at half speed. Changes in the rate of sound file playback were calculated in real-time utilizing the "groove~" object in the Max/MSP programming language.

Seismic data was downloaded as audio files from a publicly accessible website maintained by the United States Geological Survey (USGS) science program. The first example contained data from a magnitude 5.1 event that was recorded in Parkfield, California (1994). The second example contained data from a magnitude 6.5 event that was recorded in Petrolia, California (1992). Researchers from U.C. Berkeley recorded both seismograms [14].



	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Correct - Digitally Generated	6	5	6	6	6	6	6	6	4	4	6	5	6	6	6	6	2	6	6
Correct - Audification	15	15	13	13	13	11	11	11	12	12	10	10	7	5	4	3	6	2	2

Figure 2. Individual performance on the identification task sorted by number of correct responses (highest to lowest). This stacked bar graph provides the number of correctly identified audification examples (bottom) as well as the number of correctly identified digitally manufactured examples (top).

Two examples of audified solar wind data were created for this experiment. The first example was downloaded from the NASA's Coordinated Data Analysis Web (CDAWeb) as a comma-separated text file. This type of data, which merges records from multiple satellites, is referred to as OMNI data and is available to the general public publically available. This specific file contained solar wind hourly averaged bulk proton flow speed (km/s) measurements spanning the years 1963 to 2010 inclusively, and was 421,057 entries in length.

The second solar wind example was generated with data collected by the Solar Wind Ionic Composition Spectrometer (SWICS) instrument on the Advanced Composition Explorer (ACE) satellite. This data measured the variance of the solar magnetic field at 16-second intervals, and was gathered over the course of the year 1997. The source file was downloaded from a publically accessible data repository [15], this file was 112,104 data samples in length.

The neuronal firing data was collected from a probe during a Deep-Brain Stimulation (DBS) surgical procedure. The probe, measuring approximately 40-microns in circumference circumference, recorded micro-voltage fluctuations at a rate of 30,000 samples per second. This audio was converted to a sampling rate of 44,100 for playback within the experimental interface. The two neuronal firing examples were taken from separate files; one file measured 83.3 megabytes in size, while the other measured 1.69 gigabytes. After audifiaction, a sub-section was chosen from each file that contained prominent firings from a single-neuron (as identified by a researcher experienced in close-listening to audio from DBS procedures).

The white noise and sinusoidal listening examples were generated with the Max/MSP computer-music programming language, utilizing the "noise~" and "cyle~" objects respectively. The frequency of the sine wave example was 440hz.

2.4. Apparatus

The experiment was conducted utilizing a 15-inch MacBook Pro running the Mac OS X operating system (Version 10.6.7). All participants used Sony MDR-7509HD Dynamic stereo headphones for all listening examples. The software interface was designed and constructed within the Max/MSP computer-music programming environment (Version 5.1.8). A standalone application was created, which saved experimental data as files in ".txt" format. Before beginning the experiment, participants were prompted to provide their first name, middle initial, and last name; this data was parsed and the resulting initials were used to create unique file names.

3. RESULTS

3.1. Overview

In this forced-choice listening task with 5 possible responses, a 20% success rate across all 24 listening examples would result in an average of 4.8 correct responses. This success rate would indicate chance-performance. Results from this identification task have been summarized in **figure 2**. The average number of correct responses across all participants (and

all listening examples) was 14.68, with a standard deviation of 4.41. This finding is considered to be extremely statistically significant when compared against chance performance ($P < 0.0001$). Measures of statistical significance in all cases were calculated utilizing a t-test, unless otherwise noted the statistical mean was measured against chance performance. The highest number of correct responses was 21 (1 participant) and the lowest number of correct responses 8 (recorded by 3 participants). Information regarding the number of correct responses for each listening example has been provided in **figure 3**.

	Audio Example	Playback Speed	Correct Response
1.	Solar-Wind Data (Proton Flow Speed)	1/2	63.2%
2.	Brainwave Data (Example 1)	Original (1)	31.6%
3.	Seismic Data (Parkfield 1994, Mag 5.1)	1/2	63.2%
4.	Brainwave Data (Example 2)	3/4	47.4%
5.	Solar-Wind Data (Proton Flow Speed)	3/4	57.9%
6.	Sine Tone (440hz)	3/4	100.0%
7.	Seismic Data (Petrolia 1992, Mag 6.5)	Original (1)	26.3%
8.	Solar Wind Data 2 (magnetometer)	3/4	26.3%
9.	White Noise	3/4	84.2%
10.	Brainwave Data (Example 2)	Original (1)	63.2%
11.	Sine Tone (440hz)	1/2	94.7%
12.	Solar-Wind Data 2 (Magnetometer)	Original (1)	57.9%
13.	Seismic Data (Parkfield 1994, Mag 5.1)	3/4	63.2%
14.	Brainwave Data (Example 2)	1/2	47.4%
15.	White Noise	Original (1)	84.2%
16.	Seismic Data (Petrolia 1992, Mag 6.5)	3/4	47.4%
17.	Sine tone (440hz)	Original (1)	94.7%
18.	Brainwave Data (Example 1)	1/2	47.4%
19.	White Noise	1/2	89.5%
20.	Brainwave Data (Example 1)	3/4	42.1%
21.	Seismic Data (Petrolia 1992, Mag 6.5)	1/2	52.6%
22.	Solar Wind Data 2 (magnetometer)	1/2	73.7%
23.	Solar Wind Data (Proton Flow Speed)	Original (1)	57.9%
24.	Seismic Data (Parkfield 1994, Mag 5.1)	Original (1)	52.6%

Figure 3. Complete list of audio examples, relative playback speed, and percentage of correct responses. This is the order in which the listening examples were provided to all participants.

3.2. Correlative Evaluation

Participants completed the pre-test questionnaire, listening task, and post-test questionnaire in an average of eleven minutes. No correlation was found between above-average and below-average completion time and the number of correct responses ($P < 0.7524$). Similarly, no significant correlation was found between gender and recognition ability ($P < 0.73$). There was no significant difference between the performance of participants aged 24 and younger, or 25 and older ($P < 0.8541$). Participants with Masters or PhD degree correctly identified an average of 16 examples (out of a total 24), while participants who completed high school or a received a bachelor's degree correctly identified an average of 14.21 ($P < 0.45$). Participants with 7 or more years of musical training successfully identified an average of 17.75 examples, while participants with zero to six years of musical training identified an average of 13.87 ($P < 0.12$). These results were determined to be statistically insignificant based on the small sample size (see **figure 4**).

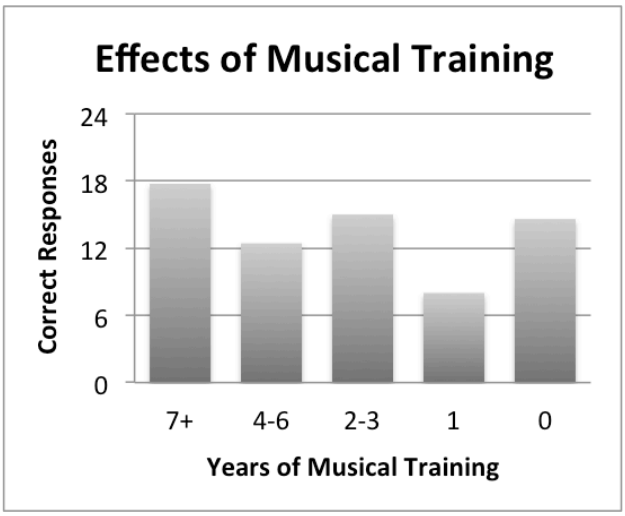


Figure 4. No statistically significant correlation was found between a participant's level of musical training and performance on the identification task.

3.3. Controlling For Pre-Exposure

The following section independently evaluates recognition ability for the six digitally generated sounds (three white noise and three sinusoidal) as opposed to the eighteen audified examples (six solar wind, six neuronal and six seismic). All participants correctly identified an average 9.21 of the 18 audified examples, which indicates a performance significantly better than chance ($P < .0001$). The 13 participants who had previously been exposed to at least one of the listening examples were able to correctly identify an average of 9.85 of the 18 audified data examples, this is considered to be extremely statistically significant when compared to chance ($P < 0.0001$). The 6 participants who had not previously been exposed to any of the listening examples were able to correctly identify an average of 7.83 of the 18 audified data sets. This performance is 23 percentage points higher than chance, however, the sample size is not large enough to determine statistical significance ($P < 0.09$).

All participants correctly identified an average of 5.47 of the 6 digitally manufactured sounds, which indicates a performance significantly better than chance ($P < .0001$). Participants who had never previously been exposed to any of the audio examples correctly identified an average of 5.67 of the 6 digitally manufactured sounds ($P < .0001$), while participants who had been previously exposed to some of the audio examples correctly identified an average of 5.38 out of 6 of digitally generated sounds ($P < .0001$). The performance difference between the two groups in the identification of the digitally generated sounds was statistically negligible ($P < 0.6$).

4. DISCUSSION

The objective of this experiment was to determine whether participants who received no training were able to identify audified data sets at a rate above chance. One notable outcomes of this experiment was that participants who had

never been exposed to audified data sets (6 of the total 19) were able to recognize the audification examples at a rate of 43.5%, which was 23 percentage points above chance. However, the sample size of individuals with no exposure was not large enough to determine statistical significance ($P < 0.0907$). A future experiment should pre-select individuals with no exposure to audified data of any kind in order to determine recognition ability for individuals with no previous exposure.

The success rate for identifying audified data sets was found to be 11% higher for participants with pre-exposure to audified data sets (54.7%) than individuals without pre-exposure (43.7%,), and this success rate was found to be statistically well above chance. This finding indicates that exposure to audified data could greatly assist in the future recognition of audified data sets, which supports the previous finding that individuals can improve recognition of non-musical auditory stimuli with training [13].

When controlling for previous exposure to any of the provided listening examples, all participants statistically performed well above chance in the recognition of white noise and sinusoidal waveforms. This indicates that participants with no previous exposure to audified data were able to discriminate between audified data and these digitally manufactured sounds without training. This provides strong support for the previous assertion that auditory data analysis can be beneficial in the identification of equipment-induced noise, particularly in the training of non-experts [2].

Many steps could have been taken to improve upon the design of this experiment. Several participants, when prompted for additional feedback in the post-test questionnaire, mentioned that they recognized repeated audio examples, despite the fact that recurring examples were always played back at different speeds. It was noted that this could be a "confounding element" as participants may try to "match… answers to the pervious ones to be as consistent as possible." As suggested by Levitin, the order of examples could have been randomized in order to minimize any bias imposed by potential "order effects" [16]. All participants correctly identified the sinusoidal waveform upon first listening, while the identification rate dropped slightly the second and third time it was presented. A randomization of presentation order across participants would be necessary in order to determine whether the playback rate of this specific sample had any impact on the number of correct responses.

One participant provided the following additional feedback: "Sometimes I wanted to put none of these I felt like the noise presented didn't sound like any of the 5 categories." This points to potential priming effects induced by the limited forced-choice selection. Participants may have responded significantly differently had they been provided with an option for "Other – This sounds like a type of audified data which is not included in this list." If the purpose of a future study were to examine the benefits of audification in exploratory data analysis, a forced choice paradigm might include an "other" option with space provided for free response. In this way the experiment could extract some ideas as to what untrained listeners believing they are hearing when they are free to craft novel responses in their own words.

In addition to these improvements, a multi-frequency auditory threshold test could have been administered to establish the presence of a healthy audiometric threshold in all participants. A single-band threshold test was not found to be sufficient in this task

5. CONCLUSION

Audification has proven successful in uncovering new insights that would otherwise be overlooked through traditional analysis methods [2, 3]. However, no methodological framework has been established for how this process may be successfully implemented for exploratory data analysis across a wide range of scientific disciplines. The objective of this experiment was to determine whether participants who received no training were able to identify audified data sets at a rate above chance in a forced-choice listening task. Participants who had never been exposed to audified data sets were able to recognize the audified examples at a rate that was 23 percentage points above chance performance; however, the sample size of individuals with no exposure was not large enough to determine statistical significance. When controlling for previous exposure to any of the provided listening examples, all participants statistically performed well above chance in the recognition of digitally generated sounds (White Noise and Sinusoidal waveforms). This indicates that participants with no previous exposure to audified data were able to discriminate between audified data and digitally generated sounds.

Upon repeated listening, pattern-recognition processes within the brain rapidly begin to enhance deeply embedded structural details of extremely noisy signals [17]. Exposure to audified data could greatly assist in the future recognition of audified data sets, which supports the previous finding that individuals can improve recognition of non-musical auditory stimuli with training [13]. A future experiment should pre-select individuals with no exposure to audified data of any kind in order to determine recognition ability for individuals with no previous exposure. Another future study should examine the benefits of audification in exploratory data analysis through a forced choice paradigm with an "other" option. This free-response space would allow participants to craft novel responses in their own words, which could provide valuable insight.

References

[1] Kramer, G., Walker, B., Bonebright, T., Cook, P., Flowers, J., Miner, N.; Neuhoff, J., Bargar, R., Barrass, S., Berger, J., Evreinov, G., Fitch, W., Gröhn, M., Handel, S., Kaper, H., Levkowitz, H., Lodha, S., Shinn-Cunningham, B., Simoni, M., Tipei, S. *Sonification report: status of the field and research agenda*. International Community for Auditory Display, 1999.

[2] Hayward, C. Listening to the Earth Sing. in Kramer, G. ed. *Auditory display: Sonification, Audification, and Auditory Interfaces*, Addison-Wesley, Reading, Mass., 1994, 369-404.

[3] AUTHOR.

[4] Eckart, C. *Principles of Underwater Sound*. Division of War Research. Sonar Data Division, United States. Office of Scientific Research and Development. National Defense Research Committee., California. University., 1946.

[5] Shannon, C.E. and Weaver, W. *The Mathematical Theory of Communication* University of Illinois Press, Urbana, 1949.

[6] Pollack, I. and Ficks, L. *Information of Elementary Multidimensional Auditory Displays*. ASA, 1954.

[7] Fidell, S. *Sensory Function in Multimodal Signal Detection*. ASA, 1970.

[8] Loveless, N.E., Brebner, J. and Hamilton, P. Bisensory presentation of information. *Psychological Bulletin*, *73* (3). 161-199.

[9] Yeung, E.S. Pattern recognition by audio representation of multivariate analytical data. *Analytical Chemistry*, *52* (7). 1120-1123.

[10] Bly, S. Presenting information in sound *Proceedings of the 1982 conference on Human factors in computing systems*, ACM, Gaithersburg, Maryland, United States, 1982, 371-375.

[11] Bly, S. Sound and computer information presentation *Other Information: Thesis. Portions of document are illegible*, 1982, 124.

[12] Minghim, R. and Forrest, A.R. An Illustrated Analysis of Sonification for Scientific Visualisation *Proceedings of the 6th conference on Visualization '95*, IEEE Computer Society, 1995, 110.

[13] Corey, J. *Audio Production and Critical Listening: Technical Ear Training*. Elsevier/Focal Press, 2010.

[14] United States Geological Survey: Listen for Fun., http://earthquake.usgs.gov/learn/listen/allsounds.php, 2011.

[15] (ASC), A.S.C. ACE SWICS-SWIMS-V3 Level 2 Data, http://www.srl.caltech.edu/ACE/ASC/level2/lvl2DATA_SWICS-SWIMS.html, 2011.

[16] Levitin, D.J. Experimental design in psychological research. in Levitin, D.J. ed. *Foundations of cognitive psychology: Core readings*, MIT Press, Cambridge, MA, 2002, 115-130.

[17] Kaernbach, C. Temporal and spectral basis of the features perceived in repeated noise. *J. Acoust. Soc. Am.* 91-97

VOICE OF SISYPHUS: AN IMAGE SONIFICATION MULTIMEDIA INSTALLATION

Ryan McGee, Joshua Dickinson, and George Legrady

Experimental Visualization Lab
Media Arts and Technology
University of California, Santa Barbara
ryan@mat.ucsb.edu, dickinson@mat.ucsb.edu, legrady@arts.ucsb.edu

ABSTRACT

Voice of Sisyphus is a multimedia installation consisting of a projection of a black and white image sonified and spatialized through a 4 channel audio system. The audio-visual composition unfolds as several regions within the image are filtered, subdivided, and repositioned over time. Unlike the spectrograph approach used by most graphical synthesis programs, our synthesis technique is derived from raster scanning of pixel data. We innovate upon previous raster scanning image to sound techniques by adding frequency domain filters, polyphony within a single image, sound spatialization, and complete external control via network messaging. We discuss the custom software used to realize the project as well as the process of composing a multimodal artwork.

1. INTRODUCTION

Voice of Sisyphus relies on an Eisensteinian process of *montage* [1], the assembling of phrases with contrasting visual and tonal qualities as a way to activate change that can be considered as a narrative unfolding. Whereas cinematic montage involves the contrast of discontinuous audio-visual sequences as a way to build complexity in meaning, in this work, the referent photograph that is processed does not ever change, except through the filtering that generates the tonal and visual changes. As the composition evolves but then returns to where it began, the event brings to mind the Greek myth of king Sisyphus who was compelled to ceaselessly roll an immense boulder up a hill, only to watch it roll back down repeatedly. The intent is to have a continuously generated visual and sound composition that will keep the spectator engaged at the perceptual, conceptual, and aesthetic levels even though the referent visual source is always present to some degree.

Voice of Sisyphus was partially inspired by the overlay of image processing techniques in Peter Greenaways 2009 film, *Wedding at Cana*[1], a multimedia installation that digitally parses details of the 1563 painting by the late Renaissance artist Pablo Veronese in a 50 minute video. The filmmaker skillfully uses computer vision techniques to highlight, isolate, and transform visual details to explore the meaning of the visual elements in the original painting.

The project evokes two early digital works by Legrady. *Noise-To-Signal*[2] (1986) is an installation artwork that uses digital processing to explore the potential of image analysis, noise, and Information Theory's definition of noise to signal. *Equivalents II*[3], realized in 1992, is another interactive digital media artwork that implements 2D midpoint fractal synthesis as a way to create organic-looking abstract images whose abstract cloud-like visual forms were defined by textual input provided by viewers. Both artworks integrated synthesis algorithms to generate cultural content through computational creation of images.

Most experiments examining the relationships between sound and image begin with sounds or music that influence the visuals. Chladni's famous 18th century "sound figures" experiment involves visual patterns generated by playing a violin bow against a plate of glass covered in sand[2]. 20th century visual music artists often worked by tediously synchronizing visuals to preexisting music. Though, in some cases, the sounds and visuals were composed together as in *Tarantella* by Mary Ellen Bute. Today, visual artists often use sound as input to produce audio-reactive visualizations of music in real-time.

Less common are technical methodologies requiring images as input to generate sound. However, in 1929 Fritz Winckel conducted an experiment in which he was able to receive and listen to television signals over a radio[2], thus resulting in an early form of image audification. Rudolph Pfenninger's *Tnende Handschrift* (Sounding Handwriting), Oskar Fischinger's *Ornament Sound Experiments*, and Norman McLaren's *Synchromy* utilized a technique of drawing on film soundtracks by hand to synthesize sounds. *Voice of Sisyphus* continues in the tradition of the aforementioned works by using visual information to produce sound.

2. SOFTWARE

Custom software was developed to realize the artist's vision of translating an image into a sonic composition. Although *Voice of Sisyphus* is based on a particular photograph, the software was designed to be used with any image. Once an image file is imported one may select any number of rectangular regions within the image as well as the entire image itself to sonify. Greyscale pixel values within a region are read into an array, filtered, output as a new image, and read as an audio wavetable. The wavetables of multiple regions are summed to produce polyphonic sound. Consideration was taken for real-time manipulation of region locations and sizes during a performance or installation without introducing unwanted audio artifacts.

[1] http://www.factum-arte.com/eng/artistas/greenaway/veronese_cana.asp

[2] http://www.mat.ucsb.edu/g.legrady/glWeb/Projects/noise/noise.html

[3] http://www.mat.ucsb.edu/g.legrady/glWeb/Projects/equivalents/Equi.html

2.1. Related Work

Realization of *Voice of Sisyphus* necessitated the development of custom software for our approach to image sonification. A vast majority of existing image sonification software uses the so-called "time-frequency" approach [3] in which an image acts as the spectrograph for a sound. These systems include Iannis Xenakis' UPIC and popular commercial software such as MetaSynth and Adobe Audition. Their shared approach considers the entire image much like a musical score where the vertical axis directly corresponds to frequency and the horizontal axis to time. Usually the image is drawn, but some software like Audition allows the use of bitmap images and considers color as the intensity of frequencies on the vertical axis. MetaSynth uses the color of drawn images to represent the stereo position of the sound. In any case, all of the aforementioned software reads images left to right at a rate corresponding to the tempo. Reading an entire image left-to-right as a means to image sonification has been termed as *scanning* by Yeo[4].

However, our approach was to focus on different regions within an image over the course of the composition. Yeo has termed this approach *probing* [4]. Thus, unlike scanning, the horizontal axis of the image is not related to time. The composer must move or *probe* different regions of the image to advance the time of the composition. We also sought a more literal translation of images to sound than the typical spectrograph scanning approach. We felt that, although novel in their own right, spectrograph scanning approaches adhere too closely to a traditional musical score. We wanted a departure from the common practice of viewing images as time-frequency planes and sought a technique to listen to variations between different regions of an image. We wanted the resulting composition to unfold like one perceives a photograph in a non-linear fashion– first noticing some region, person, or object and then shifting the focus to other objects within the scene.

To produce a literal translation of image regions to sounds we began by looking at the pixel data itself. One convenient constraint was that the image chosen by the artist for the project was black and white so we did not have to consider color in our sonification approach. We began with a straight-forward audification of the 8-bit greyscale pixel values rescaled to be floating-point audio samples. The pixel values are read via raster scanning, that is line by line, top down into a 1 dimensional array of audio samples. We were aware of similar image sonification work by Yeo and Berger [5], but only became aware of their software interface, Rasterpiece [6], after we completed *Voice of Sisyphus*. Rasterpiece allows for regions of an image to be converted to sound via raster scanning with in-between filtering, a process similar to our own which we describe in later sections of this paper. As we will also detail in later sections, our software adds a more desirable filtering technique, multiple regions within the same image, Open Sound Control[7], removal of unwanted sound artifacts when manipulating regions, and sound spatialization.

2.2. Interface

The interface has both editing and presentation modes. Editing mode displays a panel of sliders for manipulating region parameters and clearly outlines all active regions within the image with colored rectangular boxes. One may create, remove, reposition, or resize regions via the mouse. Presentation mode removes the panel of sliders and region outlines from sight, making the application suitable for an artistic installation or performance to be controlled via Open Sound Control (OSC).

Interactive sonification has been defined as "the discipline of data exploration by interactively manipulating the data's transformation into sound."[8] Our software's ability to drastically change the sound obtained from image regions through interactive manipulation of spectral amplitude thresholds and segmentation of regions into a melody of subsections (both described in section 2.3) could be classified as a form of interactive sonification. The composition process for the resulting artwork described in section 3 involved interactive adjustment of parameters within a given model defined by the composer. While the final artwork is not interactive, the process of its creation could be described as working with a model-based sonification[9], which is interactive by definition.

Figure 1: Software Interface for *Voice of Sisyphus*

2.3. Sound Synthesis from Image Data

Currently, our software only deals with 8-bit greyscale images, and any color or other format images imported to the software will first be converted to 8-bit greyscale. The synthesis algorithm begins with a back-and-forth, top-down raster scanning of the greyscale pixel values, which range from 0 to 255 (black to white respectively). Simply scaling these values to obtain floating-point audio samples in the -1.0 to 1.0 range results in harsh, noisy sounds without much variation between separate regions in most images. These initial noisy results were not at all surprising given that the greyscale variation of an arbitrary image will contain a dense, broad range of frequencies. For instance, given a picture of a landscape, analyzing variations in each pixel value over a region containing thousands of blades of grass would easily produce a noisy spectrum with no clear partials. Of course, images can be specifically produced to contain particular spectra and result in tonal sounds [5], but we were interested in exploring the sounds resulting from different regions of any arbitrary image. In our case, *Voice of Sisyphus* uses an evocative photograph of a formal gala reception, so we might ask "What does a face sound like compared to a window?" Of course, the ability to determine high-level descriptions of image regions such as a "face" or "window" is a problem of feature recognition in computer vision, but we were interested in examining the objective differences in the pixel data of a "face" or "window" rather than what sounds we might normally associate with each of those objects. So, such high-level descriptions were

not necessary. We took a spectral-based approach to analyzing and processing each region's pixel data so that we could filter image regions to produce less noisy sounds with greater distinguishability between regions.

We applied a selection of frequency domain filters to our audification of pixel data by implementing a short-time Fourier transform (STFT) for each region. The STFT is obtained by computing a fast Fourier transform (FFT) of each region at the graphics' frame rate. Each FFT gives us amplitudes and phases for frequencies contained in that region at that time. Manipulation of these amplitudes and phases allows us to control the the spectrum of the image and, therefore, the resulting sound in real-time. Zeroing the amplitudes of frequencies above or below a cutoff produces a low-pass or high-pass filter respectively, while scrambling the phases of an FFT scrambles the pixels in an image without affecting is spectrum. Our key filter was to remove all frequencies below a variable amplitude threshold, leaving only the most prominent partials present and, thus, accentuating tonal differences between regions within a single image. Implementing this threshold denoises the resulting sounds, leaving clear tones that change as the region is moved or resized. The pixel data of regions is continuously updated to show the effect of the filters so the observer is always seeing and hearing the same data. As the sound becomes clearer from the filter's removal frequencies, the image becomes blurry. An interesting conclusion from this process is that most perceptually coherent images sound like noise while perceptually clear, tonal sounds result from very abstract or blurry images. This imposed a challenge for the composer of *Voice of Sisyphus* as he describes in later sections of this paper.

To obtain the final image and sound data after applying filters in the frequency domain we compute an inverse short-time Fourier transform (ISTFT) for each region, which gives the filtered pixel values. These new values are then scaled to the range -1.0 to 1.0 and read as an audio wavetable via scanned synthesis, a technique that can be used to scan arbitrary wavetables of audio data at variable rates using interpolation[10]. A control for the scan rate of these wavetables affects the fundamental pitch of the resulting sounds. However, the perceived pitch also changes as regions are moved and resized, causing new partials appear and disappear from the spectrum.

Before computing the FFT we can also scale the pixel data to effect the brightness of the resulting image and, therefore, amplitude of the sound. A masking effect can also be applied at this point, which acts as a bit reduction to the image and sound by quantizing amplitude values. Overall, it is important to note that the software only manipulates the image data and not the audio data. Since the audio data is continually produced in the same manner (scanning the IFFT results), changes in the sound are always directly produced from changes in the image. Simply put, in *Voice of Sisyphus* we are always seeing and hearing the same data. Figure 3 summarizes the sonic effects of the image filters.

The composition dictates the rapid movement and resizing of specific regions which caused discontinuities in our wavetables, resulting in an unwanted audible popping noise. To account for the resizing of images, all resulting audio wavetables, originally a length equal to the number of pixels in an image region, are upsampled or downsampled to a fixed size before linear interpolation is used to read the table at the desired frequency. Wavetables are then cross-faded with each other at the audio buffer rate to prevent discontinuities from the dynamically changing wavetables resulting from the movement and resizing of regions. If the region's position

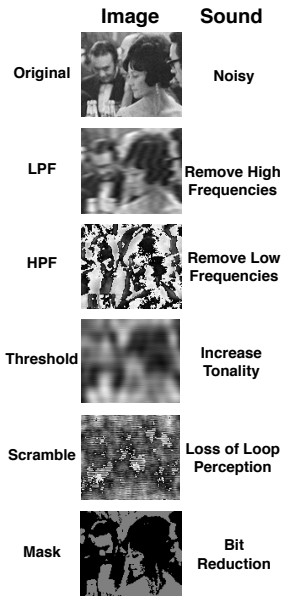


Figure 3: Effects of Image Processing on Sound

and size remain static, then the wavetable is simply looped. Scrambling the phases of the image region during the spectral processing effectively scrambles the time information of our wavetable without altering its spectra and the result is a perceptually continuous rather than looped sound.

Another challenge imposed by the composition was the desire to listen to the entire image at once. Using our STFT technique with N-point FFTs in which N is the number of pixels in the image meant taking over 1 million point FFTs at frame rate for images greater than a megapixel in size. Such computations were not suitable for the desired real-time operation. To solve this problem we added a segmentation mode for large regions which automatically subdivides them into several smaller regions of equal size. The sounds from these regions are then played-back successively left-to-right and top-down. The result of this is quite interesting–the segmentation technique is reminiscent of the step sequencers found in common electronic music hardware. Moving the segmented region produces different melodies from the resulting tones of each subsection of a region. The software's tempo slider controls the rate at which each subsection is played. Applying filters to the regions can also lead to rests in the patterns.

Regions' sounds are spatialized according to their location within the image. If a region is segmented, then the spatialization algorithm updates the position of the sound as each subsection is played, thus adding a spatial component to the aforementioned sequencer. Our method of spatialization is similar to that used in vOICe[11], an augmented reality project for the totally blind– a way to see with sound. In vOICe sounds are spatialized in 1 dimensional stereo according to their pixels' position in the horizontal image plane. *Voice of Sisyphus* uses a 2 dimensional sound plane to spatialize sounds based on to their pixels' horizontal and vertical position in the image. The installation involved a quadraphonic speaker layout, so the top left of the image was mapped to the front left speaker, the bottom left to the rear left and likewise for the right side. When more than one region is present, the

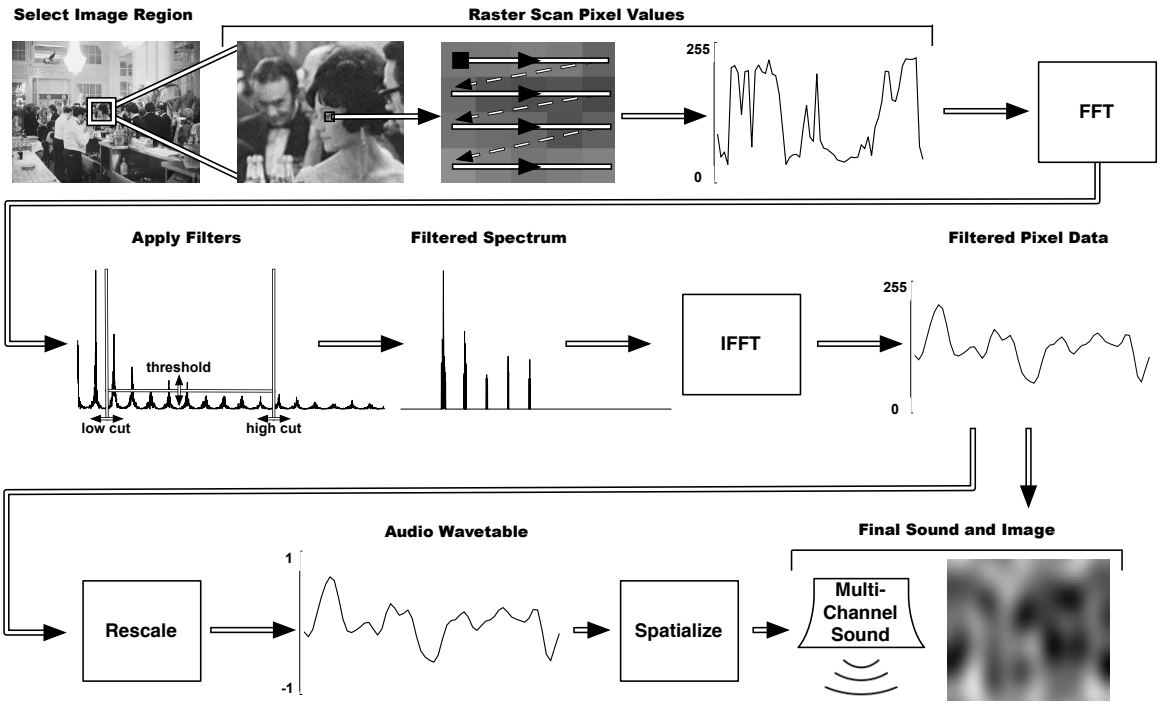


Figure 2: Sound Synthesis Algorithm for *Voice of Sisyphus*

spatialization provides a useful cue as to which sounds are coming from which regions.

The software is completely polyphonic– that is, one is free to add as many regions as desired, limited only by processor performance. Filtering and segmentation can be controlled independently for each region. Figure 4 shows an screen shot from *Voice of Sisyphus* containing 2 overlapping segmented regions producing different melodies. The screen shot also shows the software's ability to change opacity of the source image to hide or reveal it in the background.

Figure 4: Excerpt from *Voice of Sisyphus* Containing 2 Segmented Regions

2.4. Implementation

The software was programmed in C++ using OpenFrameworks[4], FFTW[5], and custom sound synthesis routines. OSC[6] allows for control of the software via any computer on the same network. In our case, the composer created a Java application with Processing that controls events in the composition. Using OSC it is also possible for many individuals to send control events to a single instance of the software, allowing a collaborative performance in which each individual controls his or her own region(s) of an image.

3. COMPOSITION

Voice of Sisyphus was conceived as an installation within a small gallery exhibition. Therefore, we wanted to compose it such that it would be compelling over a wide range of timescales. Specifically, it should have enough variation to reward prolonged engagement while also being diverse enough over the course of a couple minutes to give passing viewers a full experience.

The piece itself is an audiovisual composition that continuously cycles through a series of 8 phrases, each of which is used to convey a characteristic affect or mood. In order to prevent this repetition from becoming monotonous, the individual OSC commands or "notes" that sequence the events within each of these stages are not pre-defined but are generated in real-time based on a set of constraints. This choice ensured a theoretically limitless

[4] http://www.openframeworks.cc/
[5] http://www.fftw.org/
[6] http://opensoundcontrol.org/

number of variations over the same basic form throughout the duration that the piece was running.

Although our software supports the use of many simultaneously sonified regions, for the sake of simplicity we chose to limit ourselves to two, each of which play a distinct musical and visual role. The first region generally covers the entire width and height of our image and provides a stationary background over which moves the contrasting second region which selects smaller areas of focus. For the purposes of distinction we refer to them respectively as the "large" and "small" region.

3.1. Notation and Control

The Processing sketch that functions as our compositional score contains a series of parameters for each phrase that together determine its characteristic mood. Some values are set explicitly and remain constant though each cycle of the piece while others are chosen from within suggested constraints dynamically at the beginning of each section:

- Tempo dictates the speed at which new targets would be chosen for each parameter of the sonified regions (see quality below), as well as how quickly each region would jump about the screen if movement parameter 3 or 4 were selected. These range from 20 to 600 BPM and were set explicitly.
- Phrase lengths were chosen at random within 0.5 to 1.2 times a suggested value. The smallest suggested value was 15 seconds while the largest was 60, meaning that actual lengths range between 7.5 and 72 seconds.
- Movement type, which is the most distinguishing visual parameter, determines the way that each region moves over the image. Four main types of movement were defined, as well subtypes within each of these categories:

    1. Stationary: this was the type usually assigned to the large background region, which remains still while the small foreground region moves over it. During some sections both regions remain stationary.
    2. Smooth scanning: the region scans over the image either horizontally or vertically and either smoothly or in a randomized back-and-forth manner.
    3. Rectangular divisions: cycling randomly or in a sequential patterns in various directions, a region jumps over grid divisions of the image based upon powers of 2.
    4. Regions of interest selection: coordinates were manually gathered for all of the faces, groups of people, windows, glasses, lines, etc. within the image. These could be cycled through in various sequences.

For the regions of interest selection a "region group" variable controlled what type of object would be highlighted during each phrase or sub-phrase. For instance if *face* was selected, the smaller region would hop (on tempo) between ten pre-specified regions of the photo containing a person's face. Groups of people, windows, and glasses could each be highlighted in the same way, creating a total of 45 selectable features. The *line* setting, selects vertical strips of the image, corresponding to logical subdivisions of shapes within the scene. The ability to group visual information in this way demonstrates intent in what might otherwise appear to be a random system. Although these features are distorted and often difficult to identify, repetition suggests to the viewer semantical patterns, analogous to similar techniques used in film montage.

The fundamental frequency of both regions is set by tuning the scan rate in relation to the regions size. This tuning system provided a shorthand version of vertical harmony and values were chosen as simple ratios between the large background and small foreground region (1:2, 5:8, 10:1 etc.). Since the large region almost always remains stationary within the image and therefore has a fixed selection of pixels from which to derive frequencies, it functions like a slowly shifting drone or fixed bass over which the smaller region plays counterpoint as it moves through areas of different frequency content within the image.

Quality, which can be thought of similar to musical timbre, is actually a group of low-level filtering parameters that together determine a particular look and sound. As previously described, these include volume, mask, high-pass filter, low-pass filter, noise, and threshold. Each quality corresponds to series of suggested ranges for each of these parameters. At the beginning of a phrase, a smaller range of acceptable values is chosen from within this larger suggested range determined by the regions selected quality. While the phrase is playing, on each beat a new target is chosen from within this range of acceptable values, toward which the region interpolates. The speed at which this interpolation occurs is dictated by an independent parameter provided for the section. Large ranges of suggested values for each filter parameter are used to create phrases with a high range of timbres and quickly shifting forms, whereas a smaller range ensures that sights and sounds remain somewhat static. As a final method to ensure variation, some parameters are built in sets of 6 or 16 instead of 8, so exact repetition only occurs every 24 phrases, or 3 times through the cycle.

3.2. Compositional Themes

We wanted to depict abstracted and time-stretched methods of human/computer visual analysis. Specifically, we were interested in how an image is divided both geometrically and contextually, as well as how objects move between incoherence and recognizability. Each time our piece begins a new cycle, the entire scene is shown as a blurry and somewhat static mass of shapes. After some time the smaller region breaks off and begins to scan the image both smoothly and in jumping grid divisions. During this process, filtering of the underlying image continues to change, occasionally allowing viewers to distinguish faces and objects while at other times obscuring them completely. This is meant to mirror the way our eyes might initially try to make sense of a complicated scene. What are at first just masses of lines and shapes coalesce into identifiable forms. Likewise, as the smaller region eventually starts to directly target regions of interest within the image- faces, glasses, windows, lights, etc.- it mimics how we might scan different objects, categorizing them and placing them into logical groups based upon distinguishing features. At the climax of the piece, the entire image is shown unfiltered while the smaller region bounces as quickly as possible between important features in the image. After a few seconds this clarity fades back into a blur and the cycle starts once again.

3.3. Composition Through a Linked Audio-Visual Method

The compositional process was complicated immensely by the nature of our synthesis technique. Because sound material is generated directly by the pixels that comprise each sonified region, we

found that very interesting visuals often produced harsh or inappropriate sounds. Likewise, beautiful sonic harmonies could require very subdued or otherwise monotonous visual activity. For each section we were forced to run through countless variations and experiments in order to find parameters that produced unified and appropriate material in each sensory domain. However, as compensation for this difficulty, when the right settings are found, this method produces an audio-visual relationship that is perfectly synchronized and intuitively understood by unfamiliar audiences.

From a signal processing point of view, the results of the previous paragraph were not surprising. Non-acoustical data is inherently noisy when audified since it is not a time series of pressure data obeying the wave equation. Recognizable and meaningful visuals such as human faces are a complex arrangement of pixel values containing many frequencies when audified. The use of a variable amplitude threshold applied to the spectra of regions (outlined in section 2.3) allowed us to reduce noisy content of regions to obtain clearer, tonal sounds from otherwise complex regions. On the other hand, regions containing a simple arrangement of pixel values such architectural features (windows, edges of walls, lights, etc.), while less meaningful, lend themselves more naturally to coherent audification without heavy spectral modification.

In composing, we were not trying to substitute the visual modality of the image with a new sonic identity, but rather "add value" to the image in terms of Chion's "audio-visual contract," which describes how in film "we do not see the same thing when we also hear, and we do not hear the same thing when we also see."[12] The visual composition process of temporalizing a single image resulted in a sonic composition that in turn influenced modifications to our visual composition. The sound complemented and influenced the perception of the photograph to create an entirely new work of art. While clear portions of the image may produce otherwise unrelated abstract sounds (and vice versa), the audio-visual relationships are effective because of their precise synchronization and synthesis from the same data. We believe Chion's term "synchresis"[12] to describe a combination of *synchronism* and *synthesis* in film can also be used to describe our work.

4. INSTALLATION

Voice of Sisyphus was displayed at the Edward Cella Art+Architecture gallery in Los Angeles from November 5th, 2011 until February 4th, 2012. A single Mac Mini drove the projection and 4 channel sound for continuous operation during the aforementioned timeframe. Visitors to the gallery were free to move around the sound field or sit in a central point to experience the spatialization of the piece. After a few minutes of observation the title of the piece becomes understood. One reviewer put it, "As the image continuously reconstitutes itself and dissolves into a blurry abstraction the repetitive nature of Sisyphus' plight resonates."[7]

Most spectators quickly understood from the synchronized movement of image regions and sounds that various parts of the image were producing the audio track. However, one visitor commented on how well the chosen music matched the animation without realizing that the "music" was being generated from the image in real-time. We took this as a great complement for our sonification. Though the acoustics of the gallery space were far from ideal, the spatialization of the work proved to be quite effective as well. Listeners standing near the entrance were drawn to the center of the room once they heard the rapid movement of sounds.

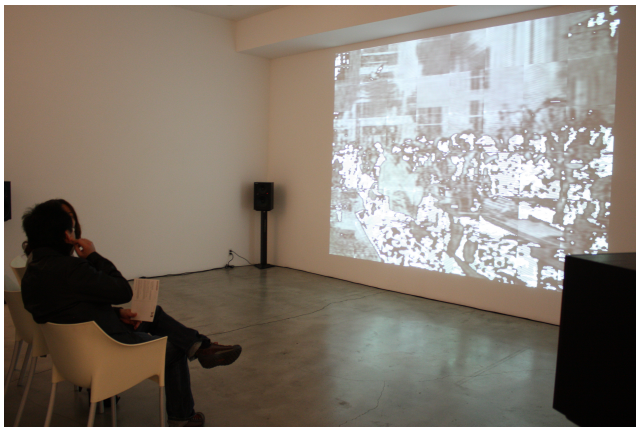


Figure 5: Installation of *Voice of Sisyphus*

5. FUTURE WORK

Future work in the area of visualization will be to implement automatic feature recognition to identify image regions of cultural interest (people, faces, etc.) using computer vision. The composition may then become autonomous from the sequential real-time sonification of image features as they are recognized by the computer. We are also interested the reverse– developing an algorithm that given a desired pitch or sound spectrum could find the best matching region within an image, thus automatically producing visuals for a composition.

6. REFERENCES

[1] B. Evans, "Foundations of a Visual Music," *Computer Music Journal*, vol. 29, no. 4, pp. 11–24, 2005.

[2] B. Schneider, "On Hearing Eyes and Seeing Ears: A Media Aesthetics of Relationships Between Sound and Image," *See this Sound: Audiovisuology 2*, pp. 174–199, 2011.

[3] C. Roads, "Graphic Sound Synthesis," *The Computer Music Tutorial*, pp. 329–334, 1996.

[4] W. S. Yeo and J. Berger, "A Framework for Designing Image Sonification Methods," in *Proceedings of International Conference on Auditory Display*, 2005.

[5] ——, "Raster Scanning : A New Approach to Image Sonification, Sound Visualization, Sound Analysis And Synthesis," in *Proceedings of the International Computer Music Conference*, 2006.

[6] ——, "Rasterpiece : a Cross-Modal Framework for Real-time Image Sonification, Sound Synthesis, and Multimedia Art," in *Proceedings of the International Computer Music Conference*, 2007.

[7] M. Wright and A. Freed, "Open Sound Control: A New Protocol for Communicating with Sound Synthesizers," in *Proceedings of International Computer Music Conference*, 1997, pp. 101–104.

[7] http://www.artillerymag.com/mini-reviews/entry.php?id=george-legrady-edward-cella-art-architecture

[8] A. Hunt and T. Hermann, "Interactive Sonification," *The Sonification Handbook*, pp. 273–296, 2011.

[9] T. Hermann, "Model-Based Sonification," *The Sonification Handbook*, pp. 399–425, 2011.

[10] B. Verplank, M. Mathews, and R. Shaw, "Scanned Synthesis," in *International Computer Music Conference*, 2000.

[11] W. Jones, "Sight for Sore Ears," *IEEE Spectrum*, February, 2004.

[12] M. Chion, W. Murch, and C. Gorbman, *Audio-Vision: Sound on Screen*. Columbia University Press, 1994.

THE SOUND OF MUSICONS: INVESTIGATING THE DESIGN OF MUSICALLY DERIVED AUDIO CUES

Ross McLachlan[1], Marilyn McGee-Lennon[2] and Stephen Brewster[2]

Glasgow Interactive Systems Group,
School of Computing Science ,
University of Glasgow,
Glasgow, G12 8QQ
r.mclachlan.1@research.gla.ac.uk[1], {first.last}@glasgow.ac.uk[2]

ABSTRACT

Musicons (brief samples of well-known music used in auditory interface design) have been shown to be memorable and easy to learn. However, little is known about what actually makes a good Musicon and how they can be created. This paper reports on an empirical user study (N=15) exploring the recognition rate and preference ratings for a set of Musicons that were created by allowing users to self-select 5 second sections from (a) a selection of their own music and (b) a set of control tracks. It was observed that sampling a 0.5 second Musicon from a 5-second musical section resulted in easily identifiable and well liked Musicons. Qualitative analysis highlighted some of the underlying properties of the musical sections that resulted in 'good' Musicons. A preliminary set of guidelines is presented that provides a greater understanding of how to create effective and identifiable Musicons for future auditory interfaces.

1. INTRODUCTION

Musicons [1] are musically-derived auditory stimuli. They are short snippets of music which can be linked meaningfully to an interface element or message (in a similar way to an Earcon for example). Musicons have so far been found to be recognisable, memorable over time and easy to learn [1]. Little is yet known, however, about what makes a 'good' Musicon. The choice of music from which to create Musicons is a key research question. Previous studies [1], [2] have shown that familiar pieces of music (such as current chart hits and musical 'memes') can be recognised from very brief samples. There has been no work investigating how recognition or preferences are affected when users themselves can select the music that the Musicons are based on.

Given that we only need a brief snippet from an entire music track to create a useful Musicon [1], [3], we need to investigate potential guidelines to aid designers in choosing the right section of the music to sample to create a useful Musicon. If the wrong section of music is selected the user may not recognise the track at all. The existing literature on Audio Thumbnailing could provide useful insights into the process of automatically extracting a representative portion of a song (such as in [4–6]). However, since different parts of a song will have different meanings to different users and since there could be more than one Musicon created from a single track, there is currently no clear way to find a Musicon algorithmically. If we could identify guidelines for Musicon creation that were able to take user's subjective preferences into account we could begin to automate the creation process based on a user's music collection.

It might also be desirable to exploit any existing relationships and emotive memories users may have with their own music tracks to enable the creation of more personalized Musicons. A Musicon personalized to a user might be more confidential to that user, easier to learn and/or remember. Understanding more about how best to create these more personalized Musicons and how well they are recognized and/or rated subjectively will provide some much needed groundwork in order that personalized Musicons can be explored in auditory interface design more thoroughly.

This paper presents a user study investigating the effectiveness of Musicons created from a user's own musical tracks. In Phase 1 of the study users were invited to upload their own music tracks and select 5 second sections based on two criteria (1) the section the user felt was most representative of the piece of music and (2) the section of the music he/she personally preferred. In Phase 2, the resulting Musicons were presented to users and evaluated in terms of both recognition and preference. Phase 3 involved analysing the resulting Musicons in terms of the underlying musical properties of the selected sections to understand better what makes a good or bad Musicon. The paper concludes with some initial guidelines for the design of successful Musicons.

2. BACKGROUND

Auditory notifications are used to alert users in a variety of applications such as calendars, social networking tools, instant messengers or SMS and telephony services. Auditory cues can take many different forms ranging from speech, to metaphorical mappings using everyday sounds (such as Auditory Icons [7]) to abstract representations with musical tones (such as Earcons [8], [9]) and super speeded-up speech (Spearcons [2]). The nature of these auditory stimuli makes each more or less appropriate depending on the user, the task and the context.

In selecting auditory cues there is an intrinsic trade-off between ease of comprehension and confidentiality; as stimuli become easier to learn they often become less private and *vice versa*. The following section briefly reviews the auditory design space with respect to ease of comprehension, confidentiality and ease of creation – all crucial factors in the design of effective and usable auditory cues. The final section evaluates Musicons in terms of each of these auditory design factors.

2.1. Comprehension

Speech messages require little or no learning if you understand the language they are spoken in. Meaningful speech messages, however, can be slower to output than other auditory cues [10]. Auditory Icons (described by Gaver [7] as "everyday sounds mapped to computer events by analogy with everyday sound-producing events") create realistic or metaphorical mappings between signifier and signified using real world sounds to represent virtual objects or actions. Since the sounds share a semantic relationship with the messages they communicate they can be easy to learn and remember. Their success, however, is fundamentally dependent on the success of the metaphor used [11] and since there is not always a mapping between a real world sound and a virtual interface action, it can be difficult to design a set of universally successful Auditory Icons. Earcons are abstract and have to be learned. They are defined by Blattner *et al.* [8] as "non-verbal audio messages used in the user-computer interface" and by Brewster *et al.* [9] as "abstract, synthetic tones that can be used in structured combinations to create audio messages". Once the association between the signifier and signified is learned, however, Earcons have been demonstrated to be a successful way to deliver auditory messages [9].

2.2. Confidentiality

Privacy or confidentiality can be an important factor in designing notifications since the messages they deliver may contain personal or sensitive information (such as with medical or personal hygiene reminders), or be delivered in a public context where only the recipient wants to intercept the message (such as when using a mobile device in a public place). Notifications that are easier to learn (such as speech or Auditory Icons) do not always offer the same level of confidentiality as more abstract auditory notifications (such as Earcons). Earcons bear no semantic relationship with the content they communicate and so those who do not know the relationship will not automatically understand the messages. Earcons (once learned) can therefore be more confidential than Auditory Icons or speech.

Spearcons are "super speeded up speech" [2] which aim to solve some of the problems associated with speech output. Text to be communicated is sped up to the point where it is not necessarily recognizable as speech yet the message can still be comprehended [2], [12]. This type of cue may provide a level of privacy not afforded with conventional speech output; if a person is not the intended recipient then the message is more difficult to intercept unintentionally. More abstract notifications can potentially offer a greater level of confidentiality since there is no semantic relationship between signifier and signified.

2.3. Creation

The key to using auditory stimuli to convey information successfully is the ability to parameterise the elements of the sound in order to encode information. With speech or Spearcons this is achieved by the concatenation of individual words in order to make sentences or structures that convey the meaning. When using speech or Spearcons, menus can be rearranged or augmented dynamically without disturbing the mapping between sounds and menu items, thus allowing interfaces to evolve without having to extend the audio design.

When using Earcons or Auditory Icons, the mapping from sound to meaning has to be created either abstractly or through a metaphor. The key difference between Earcons and Auditory Icons is the ease of parameterisation. Elements that make up an Earcon such as timbre, melody and pitch, can be extracted, analysed and manipulated using some musical skill and standard musical tools to create classes of sounds. Brewster *et al.* [13], for example, define a set of guidelines for the creation of Earcons that include recommendations of which parameters to use and how to manipulate them to maximise distinguishability. Earcons allow creation of families of sounds such that notifications and alerts that are related sound similar. Furthermore, if Earcons are designed around a grammar, a user need only learn a set of rules to understand a larger number of notifications [14].

Despite the fact that an Auditory Icon is composed of a collection of sonic elements, it is generally recorded as an atomic unit. This makes auditory icons more difficult to parameterise. There is work on the use of physical models, for example, to allow the simulation and manipulation of real-world sounds but there still remain only a small number of good models and manipulations [15]. This can make the creation of dynamic sets of Auditory Icons difficult.

In summary, there is a clear trade-off between ease of comprehension and confidentiality when using audio stimuli, one which is inherent in the difference between the abstract and metaphorical mapping of signifier to signified. Privacy issues arise with metaphorical mappings since others can potentially overhear the explicit reminders. On the other hand, the recipient may find abstract mappings more difficult to learn. The ease of creation also impacts on the usefulness of audio stimuli, since those that are easier to create make extending the audio design simpler, thus allowing the user interface to be more flexible.

2.4. Musicons

Musicons are defined as "extremely brief samples of well-known music used in auditory interface design" and have been proposed as another solution to address this gap in the audio design space [1]. By sampling a short snippet of a music track, a distinct auditory cue can be created. Musicons can enable designers to exploit existing associations and emotive memories a user may have with a piece of music to create reminders that are abstract in their relationship with the signified as well as being more memorable and potentially easier to learn.

Garzonis *et al.* [11] used pieces of music in some of their auditory icons. The BBC News and the 20th Century Fox themes were used for news and entertainment notifications, respectively. Users were able to use these effectively so this supports the notion that music may be a useful medium through which to convey information. Shellenberg *et al.* [3] asked users to identify pop tracks from short snippets of music and suggested that people could identify pieces of music well from very short snippets. McGee-Lennon *et al*. [1] created Musicons from well-known pieces of music and mapped them to everyday reminders showing that users achieved a high level of recognition (89%) sustained over a 1 week testing period.

In some respects, Musicons are comparable to Spearcons in terms of confidentiality. They can be much shorter than other types of audio stimuli and, if people do not know the association of message to musical track, the notification can provide confidentiality for the target user. Butz and Jung [16] demonstrated the use of a system that communicated notifications to a user in

musical motifs that appeared in ambient background music. Privacy was increased because the motifs used were specific to a user and would simply sound like part of the music to others. Furthermore, the notifications would not disrupt those for whom they were not intended. However, the authors concluded that the method was impractical because of the high overhead involved in composing a piece of music into which the notifications could be inserted seamlessly. The full potential of Musicons for delivery of more personalised and/or confidential messages has yet to be fully explored, though Musicons do not have as high a compositional overhead as the technique described above.

One potential advantage of Musicons over Earcons or Auditory Icons is that they could be simpler to create. A designer only needs to pick a piece of music and take a short, identifiable snippet to create a Musicon. No musical or sound design expertise is needed and there is a large amount of source material to choose from. Users could also easily create their own Musicons and they could be created automatically once Musicons are more fully understood. Schellenberg *et al.* [3] selected snippets to be "maximally representative" of the track based on the experimenter's judgment. However, except that snippets were selected to start on the downbeat at the beginning of a bar, no other guidelines were given for suitable sections from a musical track that we could use to create Musicons.

Previous work on audio thumbnailing could provide a useful insight into the creation of Musicons. An audio thumbnail is a short, representative sample of a piece of music used as a preview in order to aid search and retrieval of music tracks from a large collection [17]. However, such methods only aim to create one representative thumbnail per track [4], [6] which would be used by all users. Since we are interested in exploiting existing personal relationships and emotive memories users may have with their own music tracks, we need to investigate more subjective assessment of representativeness, a question which we address in this paper.

3. MUSICON EXPERIMENT - OVERVIEW

Previous studies have shown that pieces of music can be recognized from snippets as short as 0.2 seconds in length [1], [3]. Very little is known, however, about what makes a snippet good or bad for use as a Musicon. It is not clear *how* to pick the particular section of the music track from which to create the Musicon in terms of either performance (recognition and memorability) or preference (how pleasant it sounds).

The selection of the right section of the music to use for creating Musicons is potentially highly subjective. There is no universal metric to define 'representativeness' in terms of a section of a piece of music. We cannot assume that a universal set of Musicons is possible or ideal, and so it is necessary to test performance and preference for Musicons generated from music selected by users themselves from their own music collections.

In Phase 1 of a three part study we asked users to bring 5 music tracks from their own private collection for use in generating personalised Musicons. In Phase 2, recognition performance and preference for the Musicons were investigated. In Phase 3 we explored the underlying properties of good and bad Musicons. The following section will present each phase of the study in turn and then discuss how our findings might be used to offer initial guidelines for the design of good Musicons.

4. PHASE 1 – MUSICON CREATION

To investigate the most salient and useful features of musical tracks from which to create Musicons, an example set of Musicons was required. Results from [1] and [3] suggested that people can identify well-known tracks from very short snippets chosen by experts but there have been no studies investigating how well users can recognise snippets from tracks they have chosen themselves. To investigate this, participants were asked to supply tracks from their own music library from which a number of Musicons could be generated.

The same fifteen participants took part in both Phase 1 and 2. There were 6 females and 9 males, aged 19 - 53, none of whom reported any hearing problems. Nine of the participants reported having had formal musical training (two had a degree in music and 7 had some private tuition or training during secondary school). The remainder had no musical training.

Participants were asked to supply 5 tracks from their own music library - *Participant Tracks.* In addition, 5 *Control Tracks* were used to create Musicons that were the same across all participants. The Control Tracks, which included those used in [12], were:

- The Rembrants: I'll be there for you (Friends TV show theme)
- Ray Parker Jr: Ghostbusters
- Johan Pachelbel: Canon
- John Williams: Theme from Jurassic Park
- Theme from James Bond

These tracks were chosen because they had strong thematic associations with popular culture for the sample group of westernized adults living in the UK and the first four had proved to be effective in a previous study of Musicons [1].

By including both control and participant supplied music, the effect of **Track Type** (*Participant* vs. *Control*) on Musicon recognition and preference could be studied in Phase 2. Each participant was also asked to choose two '**selections**' from each musical track (both their own tracks and the Control Tracks). The first task was to select the section that was their personal favourite part of the track (*Favourite*). The second was to select the section they felt was most representative of the track in general (*Essence)*. Participants were asked to choose both *Favourite* and *Essence* to help us understand the different motivations behind the selection of the portion of music users might want to use for creating a Musicon from a known piece of music. Participants choose these sections on their own, using custom software. For each track, the software presented two slider bars (the knob on which corresponded to a five second slice of the song), one for 'Favourite' and one for an Essence section). Participants could adjust the sliders and play the selected clips until they were happy with their choices. Once they confirmed their selections, the software moved onto the next track. The order in which tracks were presented to participants was randomised. It was entirely possible that these two selections would overlap, or indeed be exactly the same. This, if it turned out to be the case, would in itself provide useful information.

Each of the sections selected were 5 seconds long. The decision to choose this length was made to balance the trade-off between how easy the task would be for participants and how much music there would be from which to generate Musicons. Choosing shorter selections could have been too difficult for

participants and having anything longer would have resulted in too much material from which to generate good Musicons.

Once participants had chosen all 20 of the five second sections (5 *Control/Favourite*, 5 *Control/Essence*, 5 *Participant/Favourite* and 5 *Participant/Essence*), six Musicons were generated from each – a *short* (0.2 second) Musicon from the start, middle and end of each section, and a *medium* (0.5 second) Musicon from the start, middle and end of each section. Two durations were used to analyse the effects of Musicon **length** on performance and preference.

The start, middle and end of the sections were used to generate a range of Musicons as we did not know where the most representative part within the section was located. Most of the songs selected by users could be described as, or as a sub-genre of, modern westernised pop or rock. Only one song was selected by more than one participant. Of all of the participant supplied tracks, there were only three fully instrumental tracks while the rest contained at least one singer. This resulted in a set of 120 Musicons for each participant, which was then evaluated with the same set of users in Phase 2. Each participant only evaluated his or her own set of 120 Musicons.

5. PHASE 2 – MUSICON RECALL TEST

The second phase of the study took the set of Musicons generated in Phase 1 and tested them with users to investigate recognition of, and preference for the set of Musicons. Phase 2 used a within-subjects design and took place during the same session as Phase 1. As introduced in Phase 1, the three Independent Variables were:

Track Type*:* whether a participant picked a piece of music from his/her own collection or whether it was from the control set (*Participant Track/Control Track*);

Selection*:* whether participants picked the section of the track as either favourite or essence (*Favourite/Essence);*

Length*:* the length of the Musicon (*0.2 s / 0.5 s).*

For each participant, Phase 1 produced 120 unique Musicons: 10 Tracks (5 Control and 5 Participant tracks) x 2 Selections (Favourite and Essence) x 2 Lengths (0.5s and 0.2s) x 3 Positions (Start, Middle and End). In Phase 2, participants were asked to listen to each of the Musicons and to identify the track from which it was created.

Musicons were presented in a randomised order. On hearing a Musicon, participants were asked to press a button on the experiment interface corresponding to the correct track. In total there were 10 buttons, one for each track in the experiment (5 control tracks, 5 participant tracks). This provided a measure of recognition performance for each Musicon. In addition, participants were asked to rate each of the Musicons in terms of preference. The three Dependent Variables measured were:

Identifiability*:* Whether or not the participant was able to correctly identify the track from which the Musicon was generated;

Number of Replays*:* Participants were allowed to replay each Musicon up to three times before submitting their answer. From this, it would be possible to investigate not only if a track could be identified but also how difficult it was to identify;

Preference*:* Participants were asked to rate each Musicon in terms of preference on a 5 point Likert scale (Strong Dislike, Dislike, Neutral, Like, Strong Like) based on whether they found the Musicon pleasant sounding.

5.1. Hypotheses

***H1*:** Recognition rate for Musicons generated from *Participant Tracks* will be greater than those from *Control Tracks.* Measured by higher number of *correctly identified tracks* and a lower *number of replays*;

H2: Participants will have a higher *preference rating* for the Musicons from *Participant Tracks* than *Control Tracks;*

H3: Recognition rate for Musicons created from *Essence Selections* will be higher than that for Musicons created from *Favourite Selections*. Measured by the number of *correctly identified tracks* and the *number of replays;*

H4: Recognition rate for the *0.5s* Musicons will be higher than for the *0.2s* ones. Measured by the number of *correctly identified tracks* and the *number of replays.*

5.2. Results – Recognition Rate

The recognition rate of each Musicon is shown in Table 1. Totals shown are out of 225 (15 participants x 5 Songs x 3 Positions (Start, Middle and End). Musicons that performed better than others were more correctly identified with a fewer number of replays.

	Length	
	0.2s	0.5s
Control Tracks		
Favourite	**73%(165)**	**94% (212)**
Essence	**78% (176)**	**89% (200)**
Participant Tracks		
Favourite	**69% (157)**	**84% (190)**
Essence	**72% (162)**	**85% (192)**

Table 1: Number of correctly identified Musicons.

5.2.1. Identifiability

A three-factor, repeated-measures ANOVA on Track Type, Selection and Length for the number of correctly identified Musicons showed a significant main effect for Track Type ($F(1,74)=5.513$, $p<0.05$) and a significant main effect for Length ($F(1,74)=81.799$, $p<0.01$). The main effect for Selection was not significant ($F(1,74)=0.148$, $p=0.70$). There were no significant interactions, Track Type x Selection ($F(1,74)=0.278$, $p=0.6$), Track Type x Length ($F(1,74)= 0.369$, $p=0.545$), Selection x Length ($F(1,74)=3.286$, $p =0.07$) and Track Type x Selection x Length ($F(1,74)= 2.426$, $p=0.124$).

The Musicons generated from the Control Tracks were correctly identified significantly more often than those generated from the Participant Tracks and the 0.5s Musicons were correctly identified significantly more often than the 0.2s Musicons. This partially rejects Hypothesis 1 and partially confirms H4. There was no evidence for H3.

5.2.2. Number of Replays

A Musicon could be replayed up to three times. Figure 1 shows the total number of replays over all participants for the whole experiment. The average number of replays per Musicon was small (M=0.51, SD=0.84), however, as can be seen in Figure 2, the total number of replays for 0.2s Musicons was higher than the total number of replays for 0.5s Musicons.

A three-factor, repeated-measures ANOVA on Track Type, Selection and Length for the number of replays showed no effect for Track Type ($F(1,224)=2.113$, $p=0.147$), providing no evidence for H1. The main effect for Length was significant ($F(1,224)=125.55$, $p<0.001$), as was the main effect for Selection ($F(1, 224)=4.40$, $p<0$ .05). There were no significant interactions (Track Type x Length $F(1,224)=0.159$, $p=0.69$, Track Type x Selection $F(1,224)=0.051$ $p=0.822$, Length x Selection $F(1,224)= 0.722$, $p=0.397$, Track Type x Selection x Length $F(1,224)=2.154$, $p=0.144$). Musicons of 0.2s (M=0.68, SD=0.93) were replayed significantly more often than those of 0.5s (M=0.29, SD=0.66), partially confirming H4. Musicons generated from *favourite* Selections (M=0.52, SD=0.86) were replayed significantly more than *essence* Selections (M=0.45, SD=0.79), partially confirming H3.

5.3. Results – Musicon Preference

Friedman's analysis of variance by ranks was used on the preference ratings. Differences across all factors were significant, $\chi2(3)=403.067$,$p<0.001$. *Post hoc* pairwise Wilcoxon tests with Bonferroni correction were carried out. A significant difference was observed between Musicon Lengths, $p<0.001$ and between Song Type, $p<0.001$. In general, participants preferred 0.5s (Median Rating = Like) Musicons over 0.2s Musicons (Median Rating = Neutral) and participants preferred Musicons created from the Participant supplied songs (Median Rating = Like) over those created from the Control songs (Median Rating = Neutral). There was no evidence to suggest that Section, either favourite or essence, had any effect on the preference ratings.

5.4. Discussion

The hypothesis that recognition rate for Musicons generated from Participant Tracks will be greater than Control Tracks (H1) was not supported. The Control Tracks used in this study were chosen because they had strong thematic associations with popular culture for the participant group and the results confirm that this assumption was true. The accuracy for the Participant Tracks was 78% overall, which is good, but not as high as the rates observed for the Control Tracks in this experiment (83%) and in [1] (89%). This suggests that there may be something inherently more 'identifiable' about the Control Tracks carefully chosen by experts, or that participants were more able to pick easily identifiable sections from the Control Tracks.

The hypothesis that participants would perform better with 0.5s Musicons than with the 0.2s Musicons (H4) was supported. This also confirmed the results observed by McGee-Lennon *et al.* [1] who found the same result. That the 0.2s Musicons were replayed more than 0.5s ones suggests that participants found them more difficult to recognise and adds weight to the claim that 0.5s Musicons is the most appropriate length for a Musicon.

The hypothesis that participants would perform better with Musicons created from essence sections over favourite sections

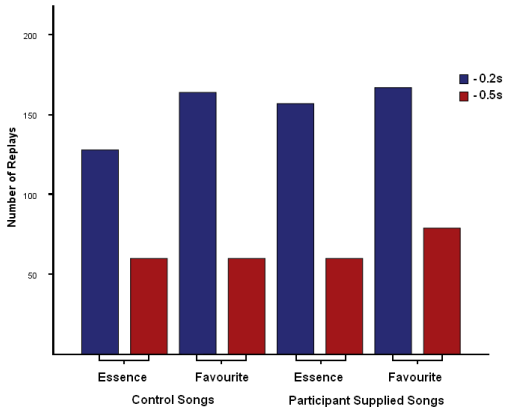


Figure 1: Overview of the number of Musicon replays.

(H3) was supported: there was no evidence to suggest that Selection had any effect on recognition rate but essence sections were replayed significantly less often than favourite ones.

The hypothesis that participants would prefer Musicons created from their own tracks over those created from Control Tracks (H2) was confirmed. The participants' responses to the Control Tracks suggested that they did not find them unpleasant but that they simply did not feel strongly either way.

6. PHASE 3 – MUSICAL SECTION ANALYSIS

The results presented above do not reveal anything about the underlying nature of the 5 second sections from which the Musicons were created. In this phase, we address two questions: (1) what are the key properties of the sections that were chosen in Phase 1? and (2) are there any similarities between the sections? We performed a qualitative analysis in which we looked at where the 5 second sections chosen in Phase 1 occurred within whole track and what musical content they contained to understand if knowledge of the properties of the music within the section may contribute to the design of good Musicons.

The analysis was designed to identify the similarities between the musical sections chosen by participants. If we could spot features that were common across well liked and easily identifiable Musicons it might help in choosing the right parts of *any* given piece of music on which to base a Musicon. The qualitative analysis involved the experimenter listening to the sections several times and looking at the underlying musical properties of the sounds to identify common compositional features between the different favourite and essence sections.

The study of the composition of a piece of music is well established in the area of Musical Analysis [18]. This broad discipline is interested in identifying the fundamental parameters or elements of a piece of music. Such analysis can highlight the underlying similarities or differences between two pieces, styles or historical periods of music by considering aspects such as form, structure, timbre and harmony. We used this approach in our analysis. Four main categories of labels were used to drive the analysis. These were derived by one of the researchers before the analysis began, based on standard definitions of musical terms which can be found on Oxford Music Online [19] and are now discussed in turn.

Structural Features: These are features relating to how the piece of the music is structured and, more specifically, where a

particular 5-second section falls within the structure. High-level structural features, such as introductions, verses and refrains, are examples. Such features are useful since they, if found to be relevant, would provide a pointer to a specific passage within a piece of music that shares a similar structure.

Timbral Features: The timbre of a piece of music refers to the overall sound and is normally defined as properties of the sound independent of rhythm or pitch. For the purposes of the analysis this is defined in terms of what instruments are present or absent with respect to the entire track, which will allow us to assess how 'full' or 'empty' the sound of this particular section is with respect to the rest of the track.

Melodic Features: These would describe whether the 5-second section contains any prominent melodic riffs, motifs or repeated melodic lines in the piece. These could be either instrumental or vocal.

Tonal and Rhythmic Features: These are features describing the salient tonal or rhythmic features of the sections. These could include, for instance, modulations (where the pitch of the track is changed substantially for effect), changes in tempo or prominent rhythmical patterns.

It was useful to augment each label with an indication of where the section lay within the whole track. For example, if a section was labelled 'Chorus/Refrain', it was useful to specify whether it was positioned nearer the start or end of the Chorus/Refrain. That the section was positioned to contain the *very* start of the chorus also proved salient (where *very* indicates that the section included the *absolute* starting point of the label, e.g. Chorus/Refrain, or contained the transition from the previous structural label, e.g. the transition from the Verse to the Chorus/Refrain). Subsequently, the augmentations 'Start', 'Very Start', 'End' and 'Very End' were included for each label.

The categories outlined above were used to guide the analysis, though the principles of Grounded Theory [20] were used to allow additional categories or themes to emerge from the data. The researcher listened to the 5 second sections several times and labelled each with as many of the features that were applicable. On each iteration, if it became clear that there were a number of sections with a common feature that was not currently being considered in the analysis, those sections would be labelled with that feature, and the feature would be considered for all sections on the next iteration. When no new features emerged, the analysis stopped.

6.1. Results

Each of the 5 second sections was labelled descriptively by the experimenter according to the underlying qualitative musical properties of each section. An overview of the labels and their frequencies can be found in Table 2 (labels with less than 5 occurrences have been omitted for brevity).

6.1.1. Control Tracks

The Control tracks were the same across participants (and did not come from the participant's own music collection). We were primarily interested in how to create Musicons from a user's own music collection. Therefore, the control tracks were not considered alongside the participant supplied songs in the detailed analysis. However, the Musicons generated from Control tracks were correctly identified more often than those created from the participant supplied ones, which either suggests that there may be something inherently more 'identifiable' about them, or that participants were better at picking easily identifiable sections from the Control tracks.

The 5 second sections that were chosen from the Control tracks were remarkably similar over all the participants. For example, of the 5 second sections chosen from The Rembrants 'I'll be there for you', 40% were of the main introduction guitar riff and 37% were of the section of the chorus during which the lyric 'I'll be there for you' is sung, while only 23% of the sections were chosen to be from other parts of the song. Similarly, of the 5-second sections chosen from Ray Parker Jr 'Ghostbusters', 53% of the sections were chosen from the verse (either where the vocalist begins to sing, or where the word 'Ghostbusters' is sung) and 37% of the sections were of the main instrumental riff, while only 10% of the sections were chosen from other parts of the song. The trend is similar for the James Bond Theme, though does not hold for either John Williams 'Theme from Jurassic Park' or Johan Pachelbel 'Canon'. The exact reasons for why the pattern is not repeated for these tracks is unknown, but both of these tracks do not contain vocals, are more classical in nature and do not have the same general structure as the western pop songs. It could be the case that the participants were more familiar with the Friends and Ghostbusters tracks, or with western pop/rock in general, and were subsequently able to make better selections. Although no strong conclusions can be drawn, it is still interesting to note the similarity between the sections. It suggests that if there are many people who are familiar with a particular song, they may have similar views on what is 'representative' of that song.

6.1.2. Participant Tracks

The majority of the labels emerging were structural in nature. Structural labels were useful in this context as they were able to transcend musical differences in genre, melody, rhythm, timbre and other intrinsically musical properties with which a composer makes a track unique. Structural similarities can group very disparate pieces of music and thus are useful for Musicon analysis. Since almost all of the user contributed songs were examples of modern western pop or rock, they were all structured in a similar way. Each song normally featured an introduction section, followed by one or more verses which were then followed by a chorus/refrain. Therefore, identifying which structural segment (e.g. introduction/verse/chorus) the 5 second section fell into was a useful way of identifying similarities between all of the 5 second sections. In total there were 150 sections (15 participants x (5 Favourite sections + 5 Essence sections)).

In addition to the structural labels, a number of melodic and timbral labels emerged as salient. The melodic labels generally indicated the presence of a strong or prominent melodic feature, such as a main riff (e.g. the main riff in Stevie Wonder 'Superstition' or in blink-182 'Apple Shampoo') or instrumental solo (e.g. the guitar solo in Santana 'Smooth', or the brass solo in Louis Prima 'Angelina, Zooma, Zooma'). The timbral features that emerged as salient generally distinguished between the presence or absence of vocals in the section. Of all of the participant supplied tracks, there were only three fully instrumental tracks while the rest contained at least one singer. Of the tracks with vocals, whether the participant's chose sections that featured the singer proved highly salient.

The most frequently observed property was the presence of a vocalist, observed in 73% of sections. In modern pop or rock music, the vocalist is often carrying the main melody. Thus, picking a section of the track containing the vocalist is important to picking a section that contains the main melody - a feature from which the track may be easily identified.

Label	Category	Frequency
Vocals	Timbral	109
Chorus/Refrain	Structural	48
Main Riff	Melodic	44
Instrumental	Timbral	41
Verse	Structural	36
Contains Track Title	Timbral	31
Chorus/Refrain – Very Start	Structural	31
Main Riff – Very Start	Melodic	30
First Verse	Structural	29
Introduction	Structural	25
Verse – Very Start	Structural	19
First Verse – Very Start	Structural	19
Full Instrumentation	Timbral	13
Introduction – Very Start	Structural	13
Middle 8	Structural	9
Instrumental Solo	Melodic	9
Climactic End-Section	Structural	8
Main Melodic Theme	Melodic	6
Chorus/Refrain – Very End	Structural	5

Table 2: Occurrences of labels in the analysis [19].

The next two highest ranking labels were 'Chorus/Refrain' (32% of sections) and 'Main Riff' (29%), which account for the highest ranking structural and melodic labels. Both the Chorus/Refrain and the Main Riff are also typically representative of western modern pop or rock music. There were a number of labels which appeared nearly as frequently as both 'Chorus/Refrain' and 'Main Riff'. The label 'Verse' appeared frequently (24%), as did 'Introduction' (17%).

Labels augmented with 'Very Start' also occurred frequently. If a 5 second section contained the very start of the Chorus/Refrain it was labelled with both 'Chorus/Refrain' and 'Chorus/Refrain – Very Start'. From this it was possible to analyse the proportion of 5 second sections *within* a particular label (e.g. all the 5 second section that were labelled with Chorus/Refrain) that were also labelled with an indication of position (e.g. Start, Very Start, End, Very End). As can be seen in Table 2, two of the highest ranking labels (Main Riff and Chorus/Refrain), have a high proportion of labels with an indication of position. 64% of all sections labelled with 'Chorus/Refrain' were also labelled with 'Chorus/Refrain – Very Start' and 68% of all sections labelled with 'Main Riff' were also labelled with 'Main Riff – Very Start'. This pattern continued with the labels 'Verse' (53% also have 'Verse – Very Start') and 'Introduction' (52% also have 'Introduction – Very Start'). All labels with this pattern are either Structural or Melodic in nature. The data suggest that if a melodic or structural feature is identified as highly representative of the track, it is likely that the *very start* of that melodic or structural feature is considered highly representative of the track.

The data suggest that there was a preference for sections that appeared nearer the beginning of tracks (sections between the introduction and first chorus). For example, labels such as 'Middle 8' (6% of sections), 'Climactic End-Section' (characterised as a unique section, appearing at the end of a song that is normally intense/exciting - it acts as a climax to the song) (5% of sections) and 'Outro' (1% of sections) occurred infrequently.

The sample of user supplied music in the study was almost entirely limited to western popular music. It is true that many underlying similarities were discovered in the 5 second sections chosen from these tracks; however, the presence of these similarities cannot be extended beyond this musical genre. This can be demonstrated with one track featured in the experiment: Duke Ellington and John Coltrane's 'The Feeling of Jazz'. This piece does not share many of the features with the other tracks in the study: it is not structured in the same way, nor does it contain any of the same salient features. The participant who chose this piece picked the very start of the introduction as his or her *Essence* selection and a section labelled 'Instrumental Solo', which occurred roughly half way through the track, as his or her *Favourite* selection. However, since the track is not structured in the common 'Introduction-Verse-Chorus' form of Western Pop/Rock music, it was difficult to draw strong comparisons between this track and all of the others in the experiment.

Overall, there was a great deal of similarity between the selections made by participants across the songs used in the experiment, suggesting that there may be common musical features that can be used to aid the selection of music from which Musicons that are representative of the piece can be created. The most frequently appearing label in the Musicon analysis was *Vocals*, suggesting that when selecting sections from which to make Musicons, the presence of a vocalist is a property that people consider representative.

7. MUSICON GUIDELINES

From the results of the previous phases the following guidelines for the design of Musicons can be drawn out:

Track Type: Musicons created from tracks that are both familiar to and liked by the user for whom they are intended are more likely to be *preferred* over those created from more generally well known tracks. Therefore, Musicons can be created by sampling snippets of music from tracks chosen by the end user to ensure a higher and more stable level of preference. However, this comes with a trade-off in performance – Musicons from participant supplied tracks were not identified as accurately as those from well known tracks. Future research should aim to investigate whether the trade-off in performance and preference changes over time, once the participant has become more familiar with the stimuli.

Length: Experimental evidence suggests that Musicons which are 0.5s in length are identified correctly and well liked.

Musical Properties: The presence of vocals was the most common feature selected by participants. Choosing a section with vocals is likely to give good Musicon performance if using western pop/rock music.

Start of Chorus/Refrain: It was common for users to select a passage of the track containing the very beginning of the first chorus or refrain. Therefore, Musicons should be sampled from a section of the track that contains vocals and the beginning of the first chorus or refrain, if using western popular music.

Start of any Melodic or Structural Feature: Although the chorus/refrain was the most popular passage in our study, there were others that were selected almost as often. If any melodic or structural feature is identified as highly representative of the track, it is likely that the *very start* of that melodic or structural feature is also considered highly representative of the track. Therefore, when sampling a Musicon from *any* Structural or Melodic passage, sample from the very start of that passage.

8. CONCLUSIONS

This research has demonstrated that by allowing users to self-select subjectively representative sections from their own music tracks, identifiable and well liked Musicons can be created. Furthermore, it was also observed that the self –selected sections were similar enough in their underlying musical features to allow for the possibility of automatic Musicon generation from an arbitrary piece of music.

Future work on Musicons is underway to focus on how well the above design guidelines can be used to create Musicons with performance and preference rates comparable to the ones observed here. Work is also planned to study how well Musicons scale and whether there is an upper limit on the number of Musicons a person can effectively remember.

The effectiveness of Musicons,(both performance and preference) compared to different types of audio stimuli, such as Earcons, Auditory Icons or Spearcons should be further investigated. The guidelines for the design of Musicons presented here provide a starting point for further investigation into the usefulness of Musicons as audio stimuli and deepen our understanding of their structure and basic composition and how this might be used to inform the design of novel auditory interfaces.

9. REFERENCES

[1] M. McGee-Lennon, M. Wolters, and R. McLachlan, "Name that tune: musicons as reminders in the home," in *ACM CHI 2011*, 2011.

[2] B. Walker and A. Nance, "Spearcons: Speech-based earcons improve navigation performance in auditory menus," *ICAD 2006*, pp. 63-68, 2006

[3] E. G. Schellenberg, P. Iverson, and M. C. McKinnon, "Name that tune: identifying popular recordings from brief excerpts.," *Psychonomic bulletin & review*, vol. 6, no. 4, pp. 641-6, Dec. 1999.

[4] M. A. Bartsch and G. H. Wakefield, "Audio thumbnailing of popular music using chroma-based representations," *IEEE Transactions on Multimedia*, vol. 7, no. 1, pp. 96–104, 2005.

[5] M. Levy, M. Sandier, and M. Casey, "Extraction of high-level musical structure from audio data and its application to thumbnail generation," in *ICASSP*, 2006, vol. 5, no. 1.

[6] H. Nawata, N. Kamado, H. Saruwatari, and K. Shikano, "Automatic musical thumbnailing based on audio object localization and its evaluation," in *ICASSP 2011*, 2011, no. 4, pp. 41–44.

[7] W. Gaver, "Auditory interfaces," *Handbook of human-computer interaction*, 1997.

[8] M. M. Blattner, D. A. Sumikawa, and R. M. Greenberg, "Earcons and icons: Their structure and common design principles," *Human-Computer Interaction*, vol. 4, no. 1, pp. 11–44, 1989.

[9] S. A. Brewster, P. C. Wright, and A. D. N. Edwards, "A detailed investigation into the effectiveness of earcons," in *ICAD*, 1992, vol. 18, pp. 471–498.

[10] A. Baddeley, "Human memory: Theory and practice," p. 423, 1997.

[11] S. Garzonis, C. Bevan, and E. O'Neill, "Mobile service audio notifications: Intuitive semantics and noises," in *OZCHI*, 2008, pp. 156–163.

[12] B. Walker and A. Kogan, "Spearcon performance and preference for auditory menus on a mobile phone," in *UAHCI*, 2009, pp. 445–454.

[13] S. A. Brewster, P. C. Wright, and A. D. N. Edwards, "Experimentally derived guidelines for the creation of earcons," in *Adjunct Proceedings of HCI*, 1995, vol. 95, pp. 155–159.

[14] S. A. Brewster, A. Capriotti, and C. V. Hall, "Using compound earcons to represent hierarchies," *HCI Letters*, vol. 1, pp. 6-8, 1998.

[15] D. Rocchesso, *Explorations in Sonic Interaction Design*. Logos Verlag Berlin, 2011.

[16] A. Butz and R. Jung, "Seamless user notification in ambient soundscapes," in *IUI '05*, 2005, p. 320.

[17] M. A. Bartsch and G. H. Wakefield, "To catch a chorus: Using chroma-based representations for audio thumbnailing," in *Applications of Signal Processing to Audio and Acoustics, 2001 IEEE Workshop on the*, 2001, no. October, pp. 15–18.

[18] I. Bent and A. Pople, "'Analysis.'," *Grove Music Online. Oxford Music Online*. Oxford Music Online.

[19] "Oxford Music Online." [Online]. Available: http://www.oxfordmusiconline.com. [Accessed: 22-Sep-2011].

[20] B. G. Glaser and A. L. Strauss, *The discovery of grounded theory: Strategies for qualitative research*. Aldine Publ., 1977.

AUDITORY SUPPORT FOR SITUATION AWARENESS IN VIDEO SURVEILLANCE

Benjamin Höferlin[†], *Markus Höferlin*[*], *Boris Goloubets*[†], *Gunther Heidemann*[‡], *Daniel Weiskopf*[*]

[†]Institute for Visualization and Interactive Systems, Universität Stuttgart, Germany
[*]Visualization Research Center, Universität Stuttgart, Germany
[‡]Computer Vision Group, Institute of Cognitive Science, University of Osnabrück, Germany
hoeferlin@vis.uni-stuttgart.de

ABSTRACT

We introduce a parameter mapping sonification to support situational awareness of surveillance operators during their task of monitoring video data. The presented auditory display produces a continuous ambient soundscape reflecting the changes in video data. For this purpose, we use low-level computer vision techniques, such as optical-flow extraction and background subtraction, and rely on the capabilities of the human auditory system for high-level recognition. Special focus is put on the mapping between video features and sound parameters. We optimize this mapping to provide a good interpretability of the sound pattern, as well as an aesthetic non-obtrusive sonification: precision of the conveyed information, psychoacoustic capabilities of the auditory system, and aesthetical guidelines of sound design are considered by optimally balancing the mapping parameters using gradient descent. A user study evaluates the capabilities and limitations of the presented sonification, as well as its applicability to supporting situational awareness in surveillance scenarios.

1. INTRODUCTION

The goal of video surveillance is to spot irregular, abnormal, or suspicious behavior of persons and objects to identify and prevent illegal or threatening actions. The huge increase of closed circuit television (CCTV) installations over the last decade shows that video surveillance has been recognized to be an appropriate method for crime prevention and evidence recording. Though, in contrast to the rapidly growing number of surveillance cameras, the monitoring capabilities stay far behind this development. The reasons are manifold, but a major factor is the high expense associated with human resources. The extent of the imbalance between recording and monitoring capabilities becomes obvious in the high camera-to-operator ratio. In their observation of 13 control rooms, Gill *et al.* [1] came across camera-to-operator ratios from 20:1 to 520:1. Keval [2] reports camera-to-operator ratios from 4:3 to 120:1 in his study of 14 control rooms. In addition to the large number of cameras to monitor, operators are often responsible for a wide variety of other tasks. A brief enumeration of such additional and often concurrently processed tasks includes [1, 2]:

- logging of incidents,
- preparation of working copies for evidence to the court or further investigation,
- tape management,
- communication with individuals inside and outside the control room, and

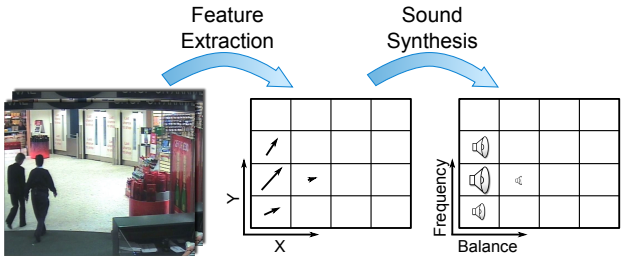


Figure 1: Segment-based feature processing and mapping to auditory parameters.

- controlling the entry/exit of the control room.

Such responsibilities lead to distraction from monitoring and hinder the detection of relevant actions and events. Further, human perception is subject to limitations that constrain the operator's event recognition ability. Such perceptual characteristics that have a strong influence on video surveillance performance include:

- the short *period of attention* when monitoring video screens (approximately 20 minutes [3]),
- difficulties to identify unexpected changes during blinks, flickers, or disruptions, called *change blindness* [4], and
- poor recognition of changes that are outside the focus of attention, termed *inattentional blindness* [5].

All these issues (mismatch of camera-to-operator ratio, additional responsibilities of CCTV operators, and perceptual constraints) point out that acceptable task performance in such high stress, multiple tasks environment requires proper situational awareness of the operators. As demonstrated by Höferlin *et al.* [6], sonification of surveillance data can support situational awareness and reduce subjective workload in multiple task scenarios.

In this paper, we apply feature extraction from video and map these features to auditory parameters (cf. Figure 1). One advantage of applying sonification to video surveillance is the complementary modality of the auditory display to the visual display, which is especially helpful when multiple target tracking and recognition tasks are performed [7]. According to the multiple resource theory, only a small degree of interferences of cognitive resources is expected in dual-task scenarios that require different mental modalities [8]. Such dual-task scenarios are typical in video surveillance [6]. Situational awareness in video surveillance further benefits from the complementary auditory display due to the excellent ability of the human auditory system to detect small changes in

sound patterns and to attract attention to those changes. As various studies pointed out (for a comprehensive overview see [9]), human auditory recognition is able to mask specific (e.g., recurrent) sound patterns from attentional processing, while being still sensitive to small variations of the sonic properties as well as to deviations to abstract rules, such as lexical, semantic, and syntactic information of human speech [10]. Such preattentive detection of change is often followed by orientation of the auditory focus of attention to the source (or auditory channel) of change. Preattentive change detection and subsequent switching of attention was well explored by magnetoencephalographical studies that explain these phenomena by differences in change-specific components of the auditory event-related brain potential, such as the *mismatch negativity* (MMN) [11].

Our approach exploits these beneficial properties of human auditory processing to support situational awareness in video surveillance. A basic assumption we make is that information relevant to surveillance monitoring is represented by changes in video signal. This means that we ascribe static parts of the video little or no relevant information. To leverage change detection capabilities of the human auditory system, our approach produces a continuous sonic pattern or soundscape of the change in video data. Further, recurrent changes in video generate an auditory texture that fades from attentional monitoring after some time of familiarization. In this state of background monitoring, sufficiently large changes of the auditory texture with respect to the familiar acoustic reference pattern reallocate attention, again. This is supported by research of the central auditory processing system that proved that MMN is only elicited after a few repetitions of a standard stimulus and only if the deviation exceeds a particular threshold [9]. Hence, we focus on the design of a non-obtrusive auditory display. Further, the parameter mapping should, to some extent, allow the interpretation of the sonification to infer from auditory display some information of the event that occurred in the monitored video. This supports a rough classification of the change recognized by the auditory signal and thus enables decision making, such as if the occurred event requires further attention by switching the visual focus to a screen.

These two main criteria for the design of our auditory display (interpretability and non-obtrusiveness) are reflected by the emphasis of this paper: the optimization between aesthetical and psychoacoustic aspects of this sonification. The goal is to find an aesthetically pleasing sonification that still conveys all of the relevant information in an interpretable manner.

1.1. Related Work

Little work has yet been published in the field of video sonification. Moreover, most of these sonifications were developed for artistic purposes (e.g., [12]) or as assistance of visually impaired people (e.g., [13]). In the context of video monitoring, we identified two related publications.

The first one is the *Cambience* system, which was developed by Diaz-Marino [14]. Besides its application in interactive arts, and as a technique to provide informal awareness between collaborators, Cambience was intended by its developer to be used as a security system that provides auditory alarms or notifications when changes occur in video. Therefore, Cambience maps video data from webcams to a sonic ecology. Differences between video frames are used to measure the level of activity in a video. Features derived from the level of activity in user-defined regions (e.g., amount of change, center of activity, and velocity) are mapped onto sound properties, such as volume, playback frequency, and stereo panning. Visual programming allows interactive definition of the mapping between sounds parameters and the features extracted from areas of interest. In the security context, Cambience provides an auditory display for process monitoring. This is closely related to the scenario we present in this paper. However, there is a distinct difference in the complexity of activities that are monitored between Cambience and the sonification presented in this paper. Cambience relies on user-defined areas of interest and is fixed on the recognition of apriori known events, such as a person entering a room. For this reason, it is constrained to be used mainly for auditory alarms. Abnormal behavior and more complex actions are thus hardly recognizable. In contrast, our approach is designed to guide attention also for apriori unknown activities and complex events that occur in the context of video surveillance.

The system by Höferlin *et al.* [6] utilizes trajectories of moving objects extracted from video data to support situational awareness of surveillance operators via a spatial auditory display. In their approach, each object trajectory is mapped to an auditory icon that moves along the object's trace in 3D sound space. By user interaction, the virtual listener's position and other parameters can be adopted to suit the monitored site. Further, the selection of auditory icons for each object class help produce a natural sound environment. The approach presented in this paper, follows a different path: one of the major differences is that we do not rely on high-level computer vision techniques, such as object tracking and classification, since these methods come with high computational cost and are not fully reliable [15]. Another difference is that we intend to avoid the mental reconstruction of the video from the auditory display. Such a translation from auditory stimulus to familiar mental representation was observed many times [16]. However, in the case of video sonification, maintenance of an imaginary video representation can be mentally demanding. We aim for a rather abstract auditory representation of relevant information and rely on the excellent capabilities of human auditory perception to detect deviations in the acoustic pattern. Although we aim for interpretability of the sonification, our primary goal is to enable auditory change detection on signal level, not on semantic level.

1.2. Contribution

According to the problem definition and related work, we aim for an auditory display meeting the following requirements:

- usage of reliable low-level computer vision features,
- comprehensive and abstract auditory display to leverage auditory change detection on signal level,
- synthesis of non-obtrusive continuous soundscape, and
- interpretability of the sonification to guide visual attention.

In the remainder of the paper, we present a novel parameter mapping sonification that copes with these requirements. This is our main contribution. As a major aspect, we tackle the often discussed issue of finding a trade-off between interpretability and aesthetics of sonification using non-linear optimization. Further, we evaluate our sonification with respect to its interpretability and support of situational awareness in video surveillance.

2. SONIFICATION DESIGN

To support the situational awareness in video surveillance, we propose a sonification system with the structure outlined in Figure 2. Besides the video display, users are provided with an auditory display based on low-level features extracted from video. These features are subsequently mapped to sonic properties of the continuous sonification. Our research prototype uses the CSound toolkit[1] for offline sound synthesis. Besides adjustment of a small set of parameters to select precision and mapping range, the auditory display does not need user intervention. Adapted values are not directly applied to the sonification, but used as input for parameter optimization to find an appropriate mapping with respect to aesthetic and psychoacoustic constraints of the auditory system.

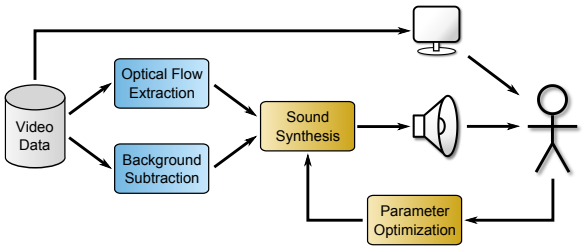


Figure 2: Data flow of the auditory display. Blue boxes depict data preparation steps by computer vision techniques. The yellow boxes represent the steps necessary for the parameter mapping sonification, described in this paper.

2.1. Data Preparation

Since we assume that only changes in video data are relevant for surveillance monitoring, we use as basic feature the dense optical flow field of two subsequent video frames. We extract the optical flow using the global method of Horn and Schunck [17]. The advantage of extracting dense optical flow over fast to compute frame differences is the availability of size and velocity information of the moving objects. For frame differencing this information is not available in the case of homogeneous colored objects, whereas the global optimization method of Horn and Schunck fills in the missing flow information by a regularization term. In addition to the motion vectors, we calculate a running average background model for foreground segmentation of the video data. This step is necessary, since optical flow calculation is prone to errors in the presence of noise and coding artifacts. Hence, motion vectors calculated in background regions are neglected for further processing. This approach helps reduce background noise and thus decrease obtrusiveness of the auditory display.

Next, we split the optical flow field into non-overlapping segments aligned in a regular grid as illustrated in Figure 1. For each segment, we calculate the average length of the contained motion vectors. This value represents the extent of activity for each segment. Please note that both the number of moving pixels and the length of the motion vectors (i.e., the velocity) influence the activity value. Hence, there are three properties for each segment to be mapped to auditory parameters: the segment's horizontal coordinate, its vertical coordinate, and its activity.

2.2. Mapping Function

There are many possible design choices for mapping the segment properties to sound parameters. However, preliminary experiments considering the users' expectations suggest the use of:

- stereo panning representing horizontal position component,
- frequency to represent the vertical component of position (rising frequency with increasing position), and
- amplitude to represent activity (low activity - soft sound).

Stereo panning and frequency dimensions are quantized, whereas amplitude is a continuous parameter. The directional information of motion in the segments is neglected. However, the direction of object movement is indirectly encoded in the temporal transition of the amplitude level between neighboring segments.

[1]CSound homepage: `http://www.csounds.com/`

From another point of view, each segment can be regarded to play its own instrument that is defined by stereo panning and frequency. If a segment shows no activity, the according instrument is muted. The complete orchestra of instruments represents the auditory display. Without aggregation to segments, motion features would be too sensitive to noise, or features of higher processing levels (e.g., trajectories) have to be used, which are prone to errors. Segmentation allows efficient sonification of low-level feature.

A key requirement of the auditory display is to convey the relevant information in an interpretable fashion. Additionally, the sonification has to be aesthetically pleasing to be non-obtrusive and broadly accepted [16]. To achieve these goals, we account for psychoacoustic aspects when defining the mapping and transfer functions. A formative user study (see Section 3) emphasized the importance of psychoacoustic aspects.

Pure tones are perceived to be unnatural, thus we use complex tones to increase natural sound sensation. For sound synthesis, each segment is represented by a periodic waveform synthesized by an additive synthesis model with 8 harmonics. Hence, the number of harmonic components we consider in the experiments is $N_H = 8$. Please note that we add only overtones that are whole multiples of the fundamental frequency in order to maintain pitch perception of complex sounds. Users can adjust the numbers of harmonics, if desired. However, although natural sounds generally have an arbitrary number of harmonics, their amplitude drops fast with higher harmonics. Thus, only few are audible and necessary for an almost natural sound sensation. By using a sine wave generator instead of MIDI sonification, we are able to tune the perceptual parameters of the sonification in much more detail, as described below. Employing the orchestral metaphor again, data features of each segment are mapped to perceptually calibrated *mini instruments* as proposed by Grond and Berger [18]. By adjusting the number of segments in each direction (horizontal and vertical), the users can trade the resolution and precision of the sonified information for the complexity of the produced soundscape.

The temporal sampling rate of the continuous sonification is set to the temporal resolution of the video data, and phases of the sine waves are adapted according to this rate to produce the impression of a continuous signal. We assume that the temporal sampling of online surveillance footage ranges from 15 fps to 30 fps. Hence, the temporal resolution of the human auditory system is capable of detecting sound changes between two successive frames. Typically, the temporal resolution for auditory change detection is beyond 20 ms, even for low frequencies (cf. [19]).

Further, we describe how we selected the transfer functions for each mapped parameter. To consider aesthetics and interpretability, we map the data properties not directly to physical sonic properties, but introduce an intermediate perceptual mapping layer.

2.3. Amplitude Mapping

To achieve linear scaling of amplitude that is necessary to interpret the information conveyed by the auditory display in the right way, we linearly map the activity value of a segment to the perceptual measure of subjective loudness S (sone at 1 kHz). Thereby, we scale the activity level to the sone interval that fits into the user-defined volume range. For the evaluation in Section 3, this range is fixed to the interval of 20 to 80 dB in the accordingly defined interval of frequency. Next step is to map loudness S to loudness level L (phon at 1 kHz) according to the non-linear relation [20]:

$$L = \begin{cases} 40 + 10\,\mathrm{ld}(S), & \text{if } S > 1 \\ 40S^{0.379}, & \text{else} \end{cases} \quad (1)$$

Finally, we map the loudness level with respect to equal-loudness-level contours to sound pressure level (dB-SPL); this value is directly fed into the CSound system and represents the amplitude of the fundamental frequency. Amplitudes of overtones are adapted accordingly and normalized by CSound. An analytical expression of equal-loudness-level contours fitted to experimental data was developed by Suzuki and Takeshima [21].

Obviously, this approach is only a rough approximation to adjust the perceived loudness of a data segment. We neglect any influence of overtones of complex sounds. Furthermore, dependencies between the complex tones of different data segments are not considered, too. A more elaborated loudness model will be considered in future work, a thorough evaluation of advanced models was presented by Skovenborg and Nielsen [22].

2.4. Stereo Panning

A segment's horizontal position component is a linearly mapped between left and right channel and scaled to fit the complete panning range. The energy of the panned signal is kept constant with the source signal. Note that we do not account for directional dependencies of loudness and pitch perception, since we expect the sonification to be used with headphones.

2.5. Frequency Mapping

To map the vertical position component of a segment to frequency, we have to consider different, sometimes opposing objectives. First, we require a linearly perceived increase of frequency for interpretability reasons; while for a pleasing sonification the tone heights of two segments should match consonant intervals. These criteria have to be met under the constraint of a limited frequency spectrum to be used. And finally, frequencies should increase monotonically with a step size of at least the perceptual just notable difference.

To find the most suitable distribution of frequencies Φ (ordered increasing set of fundamental frequencies in Hz) that copes with these competing goals, we formulate a cost function Ψ to be minimized by gradient descent in combination with simulated annealing as follows

$$\Psi(\Phi) = \gamma_l\,\Psi_l(\Phi) + \gamma_d\,\Psi_d(\Phi) + \gamma_o\,\Psi_o(\Phi) + \gamma_r\,\Psi_r(\Phi) \quad (2)$$

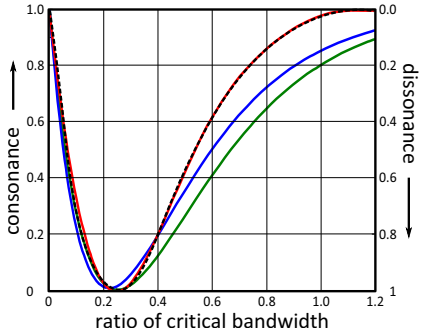


Figure 3: Perceived dissonance of pure tones as a function of the ratio of the critical bandwidth. Experimentally obtained dissonance function by Plomp and Levelt [24] (dashed line), Benson's approximation [25]: $4|x|e^{1-4|x|}$ (green), Sethares' approximation cited in [25] (blue), and our fitting in Equation 5 (red).

with γ_x being a user-defined factor to emphasize particular cost terms Ψ_x that are described in the subsections below. Note that we require the cost terms Ψ_x to be differentiable, since we use gradient descent. Further, we found that an equal distribution of the N fundamental frequencies $\varphi \in \Phi$ in the user-defined frequency range is a suitable initial value to start the gradient descent.

Linear Scaling. The first cost term Ψ_l represents the linearity of the perceived pitches: a property that is important for understanding the conveyed information. To rate the ordered set of fundamental frequencies Φ, we map each of the frequencies $\varphi_i \in \Phi$ (in Hz) to *Zwicker*'s bark scale (critical bandwidth rate, CBR), a perceptual scale of pitches that accounts for the place-spectral analysis of the cochlea [23]:

$$\mathrm{CBR}(\varphi) = 13\,\mathrm{atan}(0.00076\varphi) + 3.5\,\mathrm{atan}(\varphi/7500)^2 \quad (3)$$

As a natural measure of linearity, we take the second (smaller) eigenvalue λ_2 of the 2×2 covariance matrix of the set of vectors

$$\left\{ \begin{pmatrix} \mathrm{CBR}(\varphi_1) \\ 1 \end{pmatrix}, \begin{pmatrix} \mathrm{CBR}(\varphi_2) \\ 2 \end{pmatrix}, \ldots \begin{pmatrix} \mathrm{CBR}(\varphi_N) \\ N \end{pmatrix} \right\}$$

Therefore, we assume that at least a minimum of linearity already exists. Further, we assume the influence of sound pressure level on the perceived pitch to be already compensated by loudness-based amplitude mapping.

Consonant Intervals. To improve acceptance and reduce obtrusiveness and annoyance of our sonification, we account for aesthetics and musicality in terms of consonant intervals. Consonant complex tones exhibit harmonic vibration ratios of their partials (integer multiples) and thus sound pleasant to most people. As measure of consonance of the complex tones of the ordered set of fundamental frequencies Φ (in Hz) with N_H harmonics, we apply the method reported by Plomp and Levelt [24]. The dissonance costs Ψ_d therefore represent the sum over the degree of dissonance of two successive fundamental frequencies $\varphi_i, \varphi_{i+1} \in \Phi$ (in Hz) with their overtones:

$$\Psi_d(\Phi) = \frac{1}{N_H^2(N-1)} \sum_{i=1}^{N-1} \sum_{j,k=1}^{N_H} \mathrm{d}\left(\frac{|j\varphi_i - k\varphi_{i+1}|}{\mathrm{CB}(\sqrt{jk\varphi_i\varphi_{i+1}})} \right) \quad (4)$$

Table 1: Coefficients for sine approximation of dissonance term.

i	α	β	γ
1	2.035	4.340	-1.387
2	3.424	5.662	0.4757
3	1.680	6.469	2.873

The dissonance function d is a perceptual measure that was experimentally derived by Plomp and Levelt [24]. Although several analytical approximations have already been published, we propose a more precise fitting on sine basis (see Table 1 for coefficients, and Figure 3 for a comparison with the original data):

$$\mathrm{d}(x) = \begin{cases} \sum_{i=1}^{3} \alpha_i \sin(\beta_i x + \gamma_i) & , \text{ if } x \leq 1.2 \\ \mathrm{d}(1.2) & , \text{ else} \end{cases} \quad (5)$$

The function $\mathrm{CB}(f_c)$ provides the critical bandwidth of the center frequency $f_c = \sqrt{\varphi\widetilde{\varphi}}$ of the two compared harmonics $\varphi, \widetilde{\varphi}$ according to Zwicker and Terhardt [23]:

$$\mathrm{CB}(f_c) = 25 + 75(1 + 1.4 \cdot 10^{-6} f_c^2)^{0.69} \quad (6)$$

Finally, Ψ_d is normalized to fit the interval $[0, 1]$.

Frequency Order. It is a main requirement of our approach that frequencies in the ordered set Φ increase monotonically. Hence, we have to assure that this criterion is met for all possible solutions of the optimization. The term Ψ_o insures this by penalizing pairs of similar fundamental frequencies in Φ by the sum

$$\Psi_o(\Phi) = \frac{1}{N-1} \sum_{i=1}^{N-1} \left(\frac{0.056\,\mathrm{CB}(\varphi_i)}{\varphi_{i+1} - \varphi_i} \right)^{\alpha} \quad (7)$$

Monotonicity is enforced by the cost function approaching infinity as differences of neighbored frequencies approach zero. Each term of the sum becomes 1 if the frequency differences reach the frequency difference limen, which is about $1/18 \approx 0.056$ times the critical bandwidth [19]. The parameter $\alpha > 0$ is used to adjust the steepness of the function.

Frequency Range. The frequency range available for mapping is limited. Obviously, the human auditory system is restricted to the interval between 20 Hz and about 20 kHz. Furthermore, users may want to narrow this interval even more, for example to the range of musical pitch perception (50 Hz to 5 kHz). The cost term Ψ_r judges the fitness of Φ to match the user-defined frequency interval. Since we presume a monotonic increase in frequency (see section "Frequency Range") , we only have to compare the first and the last fundamental frequency (φ_1, φ_N) with the lower and upper frequency limits (f_l, f_u), respectively. However, we allow the range to exceed these limits at the penalty of rising Ψ_r, represented by sigmoid function terms

$$\Psi_r(\Phi) = \frac{1}{1 + e^{6 + \frac{12(f_u - \varphi_N)}{\mathrm{CB}(f_u)}}} + \frac{1}{1 + e^{6 + \frac{12(\varphi_1 - f_l)}{\mathrm{CB}(f_l)}}} \quad (8)$$

To account for different severities when exceeding the limits at different frequencies (violation of 20 Hz of a limit at 50 Hz is more severe than it is for a limit at 10 kHz), the sigmoidal cost function is scaled to the critical bandwidth (cf. Equation 6) at the particular limit frequency.

3. EVALUATION

We conducted two separate user studies to cover two different purposes. The first user study was conducted during an early stage of development and had a formative character. The goal of such formative evaluation is to provide "insight into which problems occur and why they occur", as well as to provide design feedback [26]. The second user study was designed as a validating user study and conducted in order to evaluate the effectiveness of our sonification approach. The study procedure, as well as the experimental setup, and given tasks were identical for both user studies. However, the participants and the presented auditory stimuli differed between the two user studies. Due to space constraints, we only provide a brief conclusion of the formative user study results here, and include, in exchange, a more detailed discussion on the results of the validating user study.

Experimental Setup. The experiments were conducted in a laboratory insulated from auditive distractions. The audio samples were presented with stereo headphones with volume control.

Stimuli and Tasks. The user study consisted of six sets (**S1** – **S6**) of stimuli and tasks with the purpose to answer different research questions. Auditory stimuli created from video data were presented, without showing the according videos. For **S1** – **S4**, artificial videos with moving textured hexagons were rendered (cf. Figure 4(a)). For **S5** – **S6**, surveillance footage was used (cf. Figure 4(b) and (c)). Stimuli with video data are available at our homepage[2].

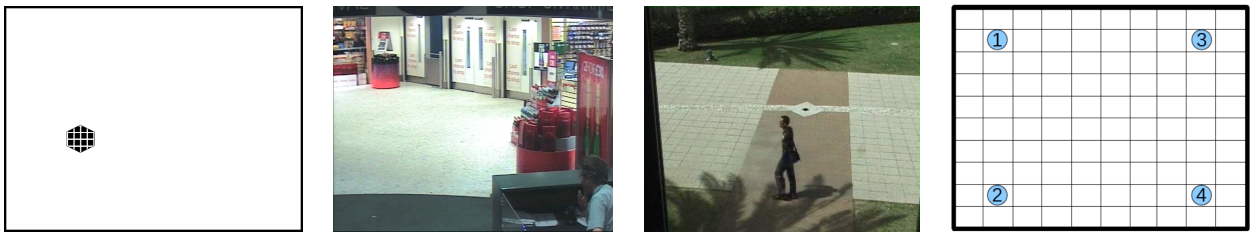
S1: ***Research Question:*** How well can object movement be detected and localized from sonification? (*Accuracy*)
Stimuli: Five stimuli, each with a single moving object. The object movement describes a rhombus, circle, two semicircles with an interruption, an eight, and a triangle.
Task: Sketching trajectories.

S2: ***Research Question:*** How well can similar object movements be distinguished? (*Discrimination*)
Stimuli: Six pairs of stimuli. Each pair consists of two objects with similar movement trajectories presented in succession. The pairs of object movements describe the following patterns: line (back and forth) – with varying slope; circle – var: radius; line (one direction) – var: acceleration; circle – var: object size; rotating object – var: object positions (long distance); rotating object – var: object positions (short distance).
Task: Sketch trajectories.

S3: ***Research Question:*** How sensitive is the sonification to distractors and noise? (*Distraction*)
Stimuli: Three stimuli, each including the movement of a single object. The applied distractors are Gaussian noise (50% normally distributed luminance changes), an image of cluttered background, and MPEG4 coding artifacts (also with cluttered background image).
Task: Sketch trajectories.

S4: ***Research Question:*** Is it possible to detect and distinguish several simultaneously occurring objects? (*Distraction*)
Stimuli: Three stimuli, showing (1) two coexistent objects,

[2] http://www.vis.uni-stuttgart.de/projekte/visual-analytics-of-video-data/sonification.html



(a) (b) (c) (d)

Figure 4: (a) Artificial video showing a hexagonal object); (b) / (c) screenshots of stimuli **S5 / S6** that were provided as context in the user study; (d) template used in the study to sketch recognized trajectories. The blue circles denote the position and order of the calibration objects in the context cue. The grid shows the granularity of the auditory display used in evaluation.

(2) three coexistent objects, and (3) two coexistent objects, where the second appears delayed.
Task: Sketch trajectories.

S5: ***Research Question:*** How well can object movement be detected and localized in real surveillance footage?
Stimuli: One stimulus based on a video from the i-LIDS multi-camera tracking scenario (duration 2:12 min). A contextual image of the video was presented along with the auditory stimuli to facilitate interpretation (cf. Figure 4(b)).
Task: Sketch trajectories.

S6: ***Research Question:*** Does the sonification allow users to detect new and abnormal patterns?
Stimuli: One stimulus based on video [27] showing a pedestrian walk (duration: 8:02 min). Additional to the sonification, a context image was provided to facilitate interpretation (c.f. Figure 4(c)). The first 1:30 min of the stimulus was provided without task in order to learn auditory patterns of normal behavior.
Task: Identification of abnormal behavior.

Study Procedure. First, subjects were asked for basic information, such as their age and profession, followed by an audiometry[3] that took about 5 min. Thereafter, they completed a PowerPoint tutorial (duration $\sim$10 min) that explained the approach and introduced the parameter mappings with the aid of artificial sample videos and their sonifications. After the tutorial, the participants were asked to answer a control question to check whether they understood the technique or not.

Then, we continued with the main evaluation that consisted of the six sets of tasks (**S1** – **S6**) and took about 40 min. Preceding to each stimulus, a *context cue* [16] was provided to enable the participants performing the interpretation tasks. The context cue was the sonification of a calibration pattern that successively showed a rotating textured object at the top left, bottom left, top right, and bottom right. After the context cue, an earcon was played that marked the beginning of the actual stimulus. For **S1** – **S4**, the participants sketched the recognized trajectories on a paper template (cf. Figure 4(d)) while the sonification was played. Acceleration/deceleration had to be marked in green, changes of the object size in red. Further, the trajectories had to be numbered according to their order of appearance. Right after each stimulus, participants had the option to correct and enhance their sketch by drawing the recognized trajectories into a second template.

For **S5**, each recognized trajectory had to be drawn on a separate template, the study operator noted the times when trajectories were identified.

For **S6**, the subjects had to verbally express recognized events. The study operator noted the events including their times.

3.1. Formative User Study

Subjects. Fifteen participants (average age 29.1 years, minimum 27 years, maximum 37 years). Sex was not considered as confounding factor for this study. Twelve participants were students or employees of our university, three participants were professional security guards. Subjects were volunteers and not paid for participation. The audiometry showed that all participants had normal hearing.

Study Results. The formative user study showed that the early version of the sonification was capable of communicating the coarse locations of the objects as well as their trajectories. The study also unveiled that aesthetics and the psychoacoustic of the sonification are critical and have to be taken into account.

3.2. Validating User Study

Subjects. Fourteen participants (average age 32.9 years, minimum 27 years, maximum 57 years). Sex was not considered as confounding factor for this study. Thirteen participants were students or employees of our university. One subject was a physician. Subjects were volunteers and not paid for participation. The audiometry showed that all participants had normal hearing.

Study Results. To judge and compare the accuracy of the sketched trajectories, we consider their start position, end position, and length. The positions are quantized on a lattice with 10 cells for each dimension (x and y, c.f. Figure 4(d)). We chose this granularity according to the expected accuracy and to limit the evaluation effort. We use the Euclidean distance between the cells of the sketched trajectory and the trajectory from ground truth (GT). The distance is normalized to [0,1] by division of the maximum cell distance (i.e., $\sqrt{10^2 + 10^2} \approx 14.14$). To compare lengths between a sketched trajectory and a GT trajectory, we count the transitions between the cells, calculate their difference, and normalize this difference by division with the GT length. A missed trajectory

[3] Applied audiometry: *HTTS-Hörtestprogramm 2.10*. URL: `http://www.sax-gmbh.de/htts/httsmain.htm`

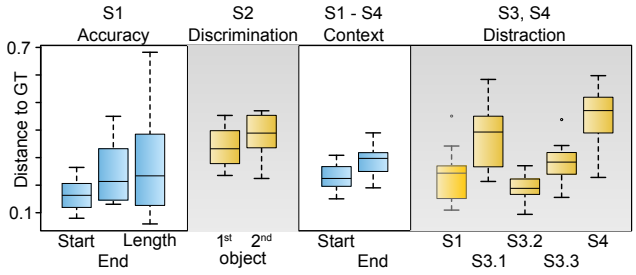


Figure 5: Boxplots of the user study results. Accuracies of the tasks are visualized as relative distances [0,1] to the ground truth. Blue boxplots represent distances of a single parameter (start position, end position, or trajectory length), while yellow boxplots show the combination of the parameters' distances. **S3.**x denotes the x^{th} stimulus of **S3**. First column: general accuracies of particular parameters; second column: accuracies at distinguishing similar object movements; third column: accuracies of start and end positions for all artificial stimuli; fourth column: sensitiveness of the accuracies with respect to distractors.

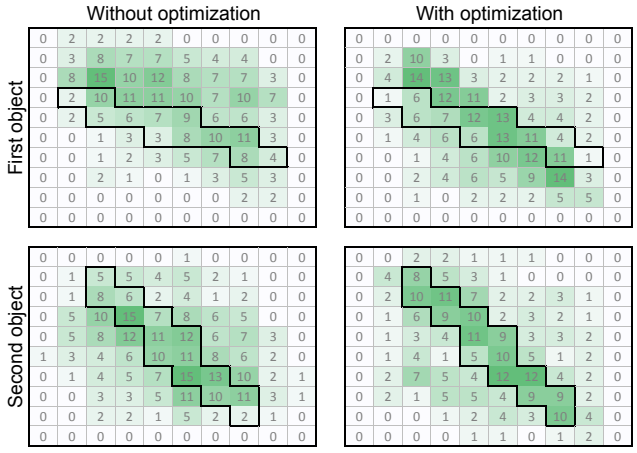


Figure 6: Example of heatmaps for the first pair of stimuli of **S2**. The frequency of how many sketched trajectories traverse a region is mapped to saturation and denoted by the numbers. The black borders denote ground truth trajectories. Left: sonification without optimization, measured during the formative study; right: proposed sonification with optimization; top: first stimulus; bottom: second stimulus with a slightly varied slope.

is penalized with the maximum distance 1 for each parameter. To summarize the accuracies, a combination of the relative distances of the parameters is calculated $\left(\frac{d_{start}+d_{end}+d_{length}}{3}\right)$.

The study results of **S1** – **S4** are depicted in Figure 5. The results of the task and stimuli set **S1** show that localization of the start (median distance: 0.16) and end position (median distance: 0.21) is possible. Moreover, the length of the trajectories can also be estimated roughly (median distance: 0.23).

The results of **S2** show that it is difficult to distinguish similar trajectories. Figure 5 shows that the combined detection accuracies of both the first (median: 0.33) and the second (median 0.39) object of the pair are worse than those of **S1** (median: 0.24). This may have two reasons: First, only a rough localization of a sonified trajectory is possible. Subjects that hear two similar trajectories focus on the movement differences and overestimate them. Second, the context cue is likely to be remembered less accurately for the second object. This is indicated by the worse results of the object appearing second. Another observation made during the study point into the same direction: the accuracy measurements of the start and end positions for all artificial video stimuli **S1** – **S4** (cf. Figure 5 (third column)) exhibit that end positions are generally detected less precisely (median: 0.30) than start positions (median: 0.22).

The localization of trajectories distracted by a background image (**S3.2**, median: 0.19) or a background image with standard MPEG4 artifacts (**S3.3**, median: 0.28) are quite robust (cf. Figure 5, median of **S1** (without distraction): 0.24). Contrary, strong noise (**S3.1**) hinders motion detection and consequently highly interferes with the sonification approach (median: 0.39). Detection of several trajectories simultaneously emerged to be most challenging: sonifying multiple trajectories at the same time drastically reduces localization accuracy (median: 0.47). While most of the subjects detected the existence of two trajectories in **S4.1** and **S4.3** (89%), it was nearly impossible to identify that there were three trajectories present in **S4.2**: only one of the fourteen participants was able to detect it. In **S4**, **S4.3** performed best (median 0.28): it is easier to localize two trajectories when they appear temporally shifted.

Figure 6 shows an example of a heatmap with the results of **S2.1** of the formative study (left) and the validating user study (right) Obviously, sonification of trajectories is more difficult to interpret, if psychoacoustic and aesthetic are not considered. As Figure 6 exhibits, perceptually correct scaling is essential to comprehend the conveyed information. Without the proposed optimization, perception among subjects seems to be more diffuse.

The results of **S5** show that it is – with some limitations – possible to detect and localize trajectories in surveillance footage. The participants were able to sketch most of the trajectories (mean: 0.79, stddev: 0.06) qualitatively correct. It is further possible to detect abnormal behavior (mean: 0.75, stddev: 0.07) due to irregularities in the auditory pattern (**S6**). Moreover, the false detection rate is quite small: on average, there was one false positive detection for each positive example in GT.

Please note that the time the subjects had to learn the standard pattern (1:30 min) as well as the time to learn the video sonification was very short. The effectiveness of the sonification can be expected to be much better when training time increases: it is likely that surveillance operators listening to the sonification for months will be able to identify smaller variations and classify them accordingly.

4. DISCUSSION AND CONCLUSION

In this paper, we introduced a sonification for video data that relies on parameter mapping of quantized optical flow fields. The sonification indicates activity in the video by an abstract sonic pattern with the aim to support situational awareness in the surveillance context. Besides this, we sketched a way to find an optimal balance between the partially opposing goals of an interpretable and aesthetically pleasing sonification. A user study showed that participants are capable of identifying abnormal events by recognizing relevant deviations of the presented soundscape. These results are a requisite to support surveillance operators and indicate that

the proposed sonification can be used as component to support situational awareness. The evaluation also exhibited the limitations of our approach, such as constraints on detection of multiple trajectories or accuracy limits for the estimation of fine movement. A consequence of these results may be the application of such sonification as supportive display.

Future work will extend the mapping by yet neglected psychoacoustic aspects, such as a more sophisticated loudness model that accounts for masking of complex tones. Besides this, optimization of other psychoacoustic aspects should be investigated, such as auditory channel separation, scalability to many displays, and change deafness.

5. ACKNOWLEDGMENTS

This work was funded by German Research Foundation (DFG) by the Priority Program "Scalable Visual Analytics" (SPP 1335).

6. REFERENCES

[1] M. Gill, A. Spriggs, J. Allen, M. Hemming, P. Jessiman, D. Kara, J. Kilworth, R. Little, and D. Swain. (2005) Control room operation: findings from control room observations. On-line Research, Development and Statistics publication. Home Office, UK. [Online]. Available: http://homeoffice.gov.uk/rds/pdfs05/rdsolr1405.pdf

[2] H. Keval, "Effective, design, configuration, and use of digital cctv," Ph.D. dissertation, University College London, 2009.

[3] M. Green, J. Reno, R. Fisher, L. Robinson, A. General, N. Brennan, D. General, J. Travis, R. Downs, and B. Modzeleski, "The appropriate and effective use of security technologies in US schools: A guide for schools and law enforcement agencies series: Research report," National Institute of Justice, Tech. Rep., 1999.

[4] R. Rensink, J. O'Regan, and J. Clark, "To see or not to see: The need for attention to perceive changes in scenes," *Psychological Science*, vol. 8, no. 5, pp. 368–373, 1997.

[5] A. Mack, "Inattentional blindness: Looking without seeing," *Current Directions in Psychological Science*, vol. 12, no. 5, pp. 180–184, 2003. [Online]. Available: http://www.jstor.org/stable/20182872

[6] B. Höferlin, M. Höferlin, M. Raschke, G. Heidemann, and D. Weiskopf, "Interactive auditory display to support situational awareness in video surveillance," in *In Proceedings of the International Conference on Auditory Display*, 2011.

[7] C. Nehme and M. Cummings, "Audio decision support for supervisory control of unmanned vehicles," MIT Humans and Automation Laboratory, Cambridge, MA, Tech. Rep. HAL2006-06, 2006.

[8] D. Boles, "Multiple resources," *International Encyclopedia of Ergonomics and Human Factors*, pp. 271–275, 2001, in: Ed. Waldemar Karwowski; Taylor and Francis, London.

[9] R. Näätänen, P. Paavilainen, T. Rinne, and K. Alho, "The mismatch negativity (MMN) in basic research of central auditory processing: A review," *Clinical Neurophysiology*, vol. 118, no. 12, pp. 2544–2590, 2007.

[10] F. Pulvermüller and Y. Shtyrov, "Language outside the focus of attention: The mismatch negativity as a tool for studying higher cognitive processes," *Progress in Neurobiology*, vol. 79, no. 1, pp. 49–71, 2006.

[11] A. Johnson and R. Proctor, *Attention: Theory and Practice*. Thousand Oaks, CA: Sage Publications, Inc, 2004.

[12] J. Pelletier, "Sonified motion flow fields as a means of musical expression," in *Proceedings of the 2008 International Conference on New Interfaces For Musical Expression*, 2008, pp. 158–163.

[13] P. Meijer, "An experimental system for auditory image representations," *IEEE Transactions on Biomedical Engineering*, vol. 39, no. 2, pp. 112–121, 1992.

[14] R. Diaz-Marino, "A visual programming language for live video sonification," Master's thesis, University of Calgary, 2008.

[15] A. Dick and M. Brooks, "Issues in automated visual surveillance," in *Proceeding of VIIth Digital Image Computing: Technique and Applications*, 2003, pp. 195–204.

[16] B. N. Walker and M. A. Nees, "Theory of sonification," in *The Sonification Handbook*, T. Hermann, A. Hunt, and J. G. Neuhoff, Eds. Logos Publishing House, Berlin, 2011, pp. 9–39.

[17] B. Horn and B. Schunck, "Determining optical flow," *Computer Vision*, vol. 17, pp. 185–203, 1981.

[18] F. Grond and J. Berger, "Parameter mapping sonification," in *The Sonification Handbook*, T. Hermann, A. Hunt, and J. G. Neuhoff, Eds. Logos Publishing House, Berlin, 2011, pp. 363–397.

[19] B. C. J. Moore, "Psychoacoustics," in *Springer Handbook of Acoustics*, T. D. Rossing, Ed. Springer Science+Business Media, LLC, New York, 2007, pp. 459–501.

[20] R. Bladon and B. Lindblom, "Modeling the judgment of vowel quality differences," *Journal of the Acoustical Society of America*, vol. 69, no. 5, pp. 1414–1422, 1981.

[21] Y. Suzuki and H. Takeshima, "Equal-loudness-level contours for pure tones," *The Journal of the Acoustical Society of America*, vol. 116, no. 2, pp. 918–933, 2004.

[22] E. Skovenborg and S. Nielsen, "Evaluation of different loudness models with music and speech material," Audio Engineering Society, Tech. Rep., 2004.

[23] E. Zwicker and E. Terhardt, "Analytical expressions for critical-band rate and critical bandwidth as a function of frequency," *The Journal of the Acoustical Society of America*, vol. 68, no. 5, pp. 1523–1525, 1980.

[24] R. Plomp and W. Levelt, "Tonal consonance and critical bandwidth," *Journal of the Acoustical Society of America*, vol. 38, no. 4, pp. 548–560, 1965.

[25] D. Benson, *Music: A Mathematical Offering*. New York, USA: Cambridge University Press, 2006.

[26] K. Andrews, "Evaluation comes in many guises," in *AVI Workshop on BEyond time and errors (BELIV) Position Paper*, 2008.

[27] N. Kiryati, T. Raviv, Y. Ivanchenko, and S. Rochel, "Real-time abnormal motion detection in surveillance video," in *19th International Conference on Pattern Recognition (ICPR)*, 2008, pp. 1 – 4.

CROSS-MODAL COLLABORATIVE INTERACTION BETWEEN VISUALLY-IMPAIRED AND SIGHTED USERS IN THE WORKPLACE

Oussama Metatla, Nick Bryan-Kinns

Queen Mary University of London
School of Electronic Engineering
& Computer Science
Mile End Road, London, E1 4NS, UK.
{oussama, nickbk}@eecs.qmul.ac.uk

Tony Stockman, Fiore Martin

Queen Mary University of London
School of Electronic Engineering
& Computer Science
Mile End Road, London, E1 4NS, UK.
{tonys, fiore}@eecs.qmul.ac.uk

ABSTRACT

We present a detailed description of the design and integration of auditory and haptic displays in a collaborative diagram editing tool to allow simultaneous visual and non-visual interaction. The tool was deployed in various workplaces where visually-impaired and sighted coworkers access and edit diagrams as part of their daily jobs. We use our initial observations and analyses of the recorded interactions to outline preliminary design recommendations for supporting cross-modal collaboration in the workplace.

1. INTRODUCTION

Every day our brains receive and combine information from different senses to understand our environment. For instance when we both see and hear someone speaking we associate the words spoken with the speaker. The process of coordinating information received through multiple senses is fundamental to human perception and is known as cross-modal interaction [1]. In the design of interactive systems, the phrase cross-modal interaction has also been used to refer to situations where individuals interact with each other while accessing a shared interactive space through different senses (e.g. [2, 3]). Technological developments mean that it is increasingly feasible to support cross-modal interaction in a range of devices and environments. But there are no practical examples where auditory displays are used to support users when collaborating with coworkers who employ other modes of interaction.

We are interested in exploring the potential of using auditory display in cross-modal interaction to improve the accessibility of collaborative activities involving the use of diagrams. Diagrams are a key form of representation often becoming common standards for expressing specialised aspects of a particular discipline (e.g. meteorologists use weather maps, architects use floor plans). However, there is currently no practical way for visually-impaired co-workers to view, let alone collaborate with their colleagues on diagrams. This is a major barrier to workplace collaboration that contributes to the exclusion and disengagement of visually-impaired individuals. Indeed, the Royal National Institute of Blind People (RNIB) estimates that 66% of blind and partially sighted people in the UK are currently unemployed [4]. Addressing the challenge of designing support for cross-modal collaboration in the workplace has thus the potential to significantly improve the working lives and inclusion of perceptually impaired workers.

2. BACKGROUND

2.1. Non-visual Interaction with Diagrams

Interest in supporting non-visual access to visually represented information grew in parallel with early developments in auditory display research [5]. A major drive of such endeavours has been and still is the potential to support individuals with temporary or permanent perceptual impairments. For example, a sonification technique pioneered in [6] displayed a line graph in audio by mapping its y-values to the pitch of an acoustic tone and its x-values to time. This sonification technique allows visually-impaired individuals to examine data presented in line graphs and tables. Current approaches to supporting non-visual interaction with visual displays employ one or a combination of two distinct models of representation; *Spatial* or *Hierarchical*. The two models differ in the degree to which they maintain the original representation when translating its visual content [7], and hence produce dramatically different non-visual interactive displays.

2.1.1. Spatial Models

A spatial model allows non-visual access to a visual display by capturing the spatial properties of its content, such as layout, form and arrangements. These are preserved and projected over a physical or a virtual space so that they could be accessed through alternative modalities. Because audio has limited spatial resolution [8], spatial models typically combine the haptic and audio modalities to support interaction. The GUIB project [9] is one of the early prototypes that employed a spatial model of representation to support non-visual interaction with a visual display. The prototype combines braille displays, a touch sensitive tablet and loudspeakers to allow blind users to interact with MS Windows and X Windows graphical environments. More recent solutions adopting the spatial model of representation typically use tablet PC interfaces or tactile pads as a 2D projection space where captured elements of a visual display are laid out in a similar way to their original arrangements. Other solutions use force feedback devices as a controller. In such instances, the components of a visual display are spatially arranged on a virtual rather than a physical plane, and can thus be explored and probed using a haptic device such as a PHANTOM Omni device [1]. The advantage of using a virtual display lies in the ability to add further haptic representational dimensions to the captured information, such as texture and stiffness, which can enhance

[1] Sensable Technologies, http://www.sensable.com

the representation of data. The virtual haptic display can also be augmented and modulated with auditory cues to further enhance the interactive experience [10, 11].

2.1.2. Hierarchical Models

A hierarchical model, on the other hand, preserves the semantic properties of visual displays and presents them by ordering their contents in terms of groupings and parent-child relationships. Many auditory interfaces are based on such a model as they inherently lend themselves to hierarchical organisation. For instance, phone-based interfaces support interaction by presenting the user with embedded choices [12]. Audio is therefore the typical candidate modality for non-visual interaction with visual displays when using hierarchies. One of the early examples that used a hierarchical model to translate visual displays into a non-visually accessible representation is the Mercator project [13]. Like the GUIB project, the goal of Mercator was to provide non-visual access to X Windows applications by organising the components of a graphical display based on their functional and causal properties rather than their spatial pixel-by-pixel on-screen representations. Other examples have employed a hierarchical model of representation to support non-visual interaction with technical drawing [14], relational diagrams [15] and molecular diagrams [16].

2.2. Cross-modal Collaboration

Despite significant progress in the use of audio and haptics in multimodal interaction design, research into cross-modal collaboration remains sparse. In particular, very little research has addressed the challenge of supporting collaboration between visually-impaired and sighted users. Nonetheless, initial investigations have identified a number of issues that impact the efficiency of collaboration in a multimodal interactive environment. An examination of collaboration between sighted and blind individuals on the Tower of Hanoi game [17], for instance, highlighted the importance of providing visually-impaired collaborators with a continuous display of the status of the shared game. Providing collaborators with independent views of the shared space, rather than shared cursor control, was also found to improve orientation, engagement and coordination in shared tasks [2]. A multimodal system combining two PHANTOM Omni haptic devices with speech and non-speech auditory output was used to examine collaboration between pairs of visually-impaired users [18] and showed that the use of haptic mechanisms for monitoring activities and shared audio output improves communication and promotes collaboration. Still, there are currently no studies of collaborations between visually-impaired and sighted coworkers. We therefore know little about the nature of cross-modal collaboration in the workplace and ways to support it through auditory design.

3. AUDITORY DESIGN IN A COLLABORATIVE CROSS-MODAL TOOL

To address the issues identified above we gathered requirements and feedback from potential users to inform an ongoing development process. We ran a workshop to engage with representatives from end user groups in order to encourage discussion and sharing of experiences with using diagrams in the workplace. Eight participants attended the workshop including participants from British Telecom and the Royal Bank of Scotland and representatives from the British Computer Association of the Blind and RNIB. Activities ranged from round table discussions exploring how participants encounter diagrams in their workplaces, to hands-on demonstrations of early audio and haptic prototype diagramming systems. The discussions highlighted the diversity of diagrams encountered by the participants in their daily jobs; from design diagrams for databases and networks, to business model diagrams, and organisation and flow charts. Additionally, participants discussed the various means they currently use for accessing diagrams and their limitations. Approaches included using the help of a human reader, swell paper, transcriptions and stationary-based diagrams, all of which share two main limitations; the inability to create and edit diagrams autonomously, and inefficiency of use when collaborating with sighted colleagues.

We chose to focus on nodes-and-links diagrams because they are frequently encountered in the workplace and we already have evaluated a single user version for audio-only interaction with such diagrams [19]. A set of requirements was thus drawn together from the workshop and other discussions to form the input to the iterative development process that followed in which a cross-modal collaborative tool was developed. Our tool [2] supports autonomous non-visual editing of diagrams as well as real-time collaboration. It allows simultaneous access to a shared diagram by augmenting a graphical display with non-visual auditory and haptic displays combining hierarchical and spatial models of representation. The tool supports user-defined diagram templates which allows it to accommodate various types of nodes-and-links diagrams such as organisation and flow charts, UML and database diagrams and transportation maps.

3.1. Graphical View

Figure 1 shows a screenshot of the graphical view of the tool. This view presents the user with an interface similar to typical diagram editors where a toolbar is provided containing various functions to create and edit diagram content. The user construct diagrams by using the mouse to select the desired editing function and has the ability to access and edit various object parameters such as labels, position, etc.

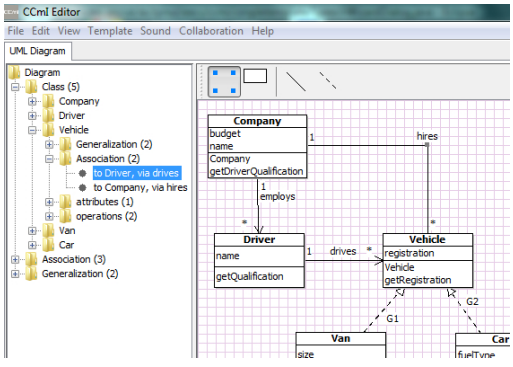


Figure 1: Graphical view (right) augmented by an auditory hierarchical view (left) embedded in the editor.

[2] An open source release of the tool and a video showcasing its features can be downloaded from: `http://ccmi.eecs.qmul.ac.uk/downloads`

3.2. Auditory Design

The design of the auditory view is based on the multiple perspective hierarchical approach described in [19]. According to this approach, a diagram can be translated from a graphical to an auditory form by extracting and structuring its content in a tree-like form such that items of a similar type are grouped together under a dedicated branch on a hierarchy. This is aimed to ease inspection, search and orientation [ibid.].

Figure 1 shows how this is achieved for a UML Class diagram. In this case, the diagram's classes – represented as rectangular shapes – are listed under the "Class" branch of the hierarchy. The information associated with each class, such as its attributes, operations and connections to other classes, is nested inside its tree node and can be accessed individually by expanding and inspecting the appropriate branches. Similarly, the diagram's associations – represented as solid arrows – are listed under the "Association" branch, and information associated with each connection can be accessed individually by inspecting its branches (see Figure 2). This allows the user to access the information encoded in a diagram from the perspectives of its "Classes", "Associations" or its "Generalisations". To inspect the content of a diagram, the user simply explores the hierarchy using the cursor keys, similar to typical file explorers, and receives auditory feedback displaying the content that they encounter.

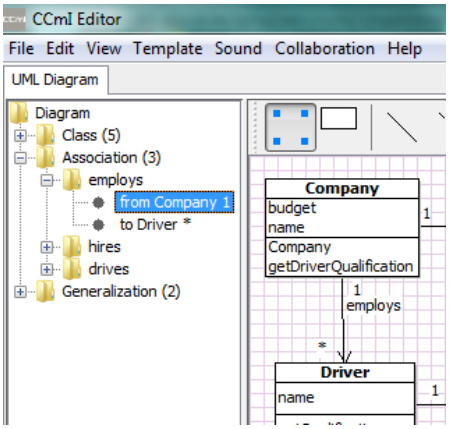


Figure 2: Hierarchical auditory view (left) where a Class diagram is accessed from the perspective of its associations.

We use a combination of speech and non-speech sounds to display encountered content. The choice of these sounds was informed through the use of an iterative prototyping approach in which candidate sounds were played to both sighted and visually-impaired users. The successful movement from one node to another is conveyed by displaying the text label of the node in speech together with a one-element earcon in the form of a single tone with a distinct timbre assigned to each type of item. This is displayed as the sequence *(earcon) + "<node name>"*. The same technique is used to highlight reaching the end or the top of a list, but in such a case a double beep tone is used instead of a single tone, and is displayed as the sequence *(earcon) + "<node name>"*, in which case the user is looped to the other end of the list. The successful expansion or collapse of a branch is also displayed using one-element earcons. An *Expand* earcon mixes frequency and amplitude modulation on a basic pulse oscillator to produce a sweep that ends with a bell like sound. A *Collapse* earcon is composed from the reversed sequence of the Expand earcon (e.g. *"Associations" + (Expand earcon)* for expanding the Associations branch, and *(Collapse earcon) + "Associations"* for collapsing it). Additionally, when a branch is expanded, a speech output is displayed to describe the number of items it contains (e.g. *"Associations" + (Expand earcon)+"three"* to convey that the diagram contains three associations). The tool allows a user to switch from one perspective on the hierarchy to another; essentially rapidly transporting to the top level of a given branch type from anywhere on the hierarchy using a single keystroke. The successful switch from one perspective to another is conveyed using a one-element earcon combined with the spoken description of the destination node. Finally, a one-element earcon is used to highlight the occurrence of illegal moves. This is referred to as the Error sound and designed as a low pitched version of the single tone browse sound. An example of an illegal move is attempting to expand an already expanded branch, or attempting to browse beyond available levels on the hierarchy.

In addition to inspecting a given diagram, the hierarchy can also be used to edit its content. To do this, the user first locates the item of interest on the hierarchy before executing a particular editing action that alters its state. For example, to remove a class from the diagram, the user would inspect the appropriate path to locate it on the hierarchy then, once found, issue the command using the keyboard to delete it. The tool then interprets the current position of the user on the hierarchy together with the issued command as one complete editing expression and executes it appropriately. The auditory hierarchical view is thoroughly described and evaluated in [15, 19].

3.3. Audio-Haptic Design

In addition to the auditory hierarchical view, we implemented a spatial model of representation to capture the layout and spatial arrangements of diagrams content. To do this, we use a PHANTOM Omni haptic device (Figure 3) to display the content of a diagram on a virtual vertical plane matching its graphical view on a computer screen (Figure 4). We designed a number of audio-haptic effects to both represent the content of a diagram and support non-visual interaction in this view.

3.3.1. Audio-Haptic Representation

The main haptic effect that we use to represent diagrams nodes and links is attraction force. Diagram nodes are rendered as magnetic points on the virtual plane such that a user manipulating the stylus of the PHANTOM device in proximity of a node is attracted to it through a simulated magnetic force. This is augmented with an auditory earcon (of a similar timbre to the one-element earcon used in the auditory view) which is triggered upon contact with the node. A similar magnetic effect is used for the links with the addition of a friction effect that simulates a different texture for solid, dotted and dashed lines. The user can thus trace the stylus across a line without deviating away to other parts of the plane while feeling the roughness of the line being traced, which increases from smooth for solid lines to medium and very rough for dotted and dashed lines respectively. Contact with links is also accompanied by one-element earcons with distinct timbres for each line style, and the labels of encountered nodes and links are also displayed in synthesised speech upon contact.

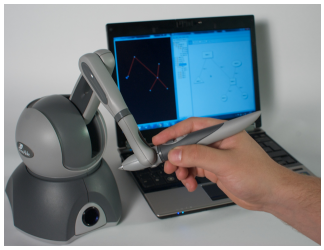


Figure 3: Interacting with the spatial haptic view using the stylus of a PHANTOM Omni haptic device.

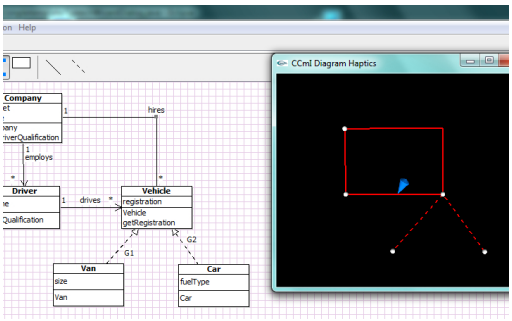


Figure 4: Spatial haptic view (right) matching the physical layout of the diagram on the graphical view.

3.3.2. Audio-Haptic Interaction

In addition to representing diagram content using various audio-haptic effects, we implemented two modes of interaction in the spatial haptic view which we refer to as *sticky* and *loose*. In a sticky mode of interaction, the magnetic attraction forces of the diagrams nodes and links are increased to make it harder for the user to snap away from contact with a given item. This simulates an impression of being "stuck" to the diagram content and thus one can trace its content by following the connections from point to point. In a loose mode of interaction on the other hand, the magnetic attraction forces are decreased such that a user can freely move around the virtual space without necessarily having to be in contact with any diagram content – in which case the haptic force is set to neutral and no auditory feedback is displayed.

The magnetic attraction forces and friction effect in the haptic view were introduced as a direct result of prototyping with visually-impaired users who found this helpful in maintaining their current position or following connections in the diagram. Similarly, the auditory feedback provided to support the haptic view was chosen to be synergistic to that in the audio-only view and was used to provide confirmatory information which was not appropriate for haptic display.

Additionally, the user has the ability to move nodes and bend links in space. This can be achieved by locating an item – or a point on a link – on the virtual plane, clicking on the stylus button to pick it up, dragging the stylus to another point on the plane, then dropping it in a new desired location with a second button click. We designed two extra features to support this drag-and-drop action. First, three distinct auditory icons are used to highlight that an item has been successfully picked up (a short sucking sound), that it is being successfully dragged in space (a continuous chain-like sound) and that it has been successfully dropped in the new location (the sound of a dart hitting a dartboard). Second, a haptic spring effect is applied, linking the current position of the stylus to the original position of where the item was picked up from. This allows the user to easily relocate the item to its original position without loosing orientation on the plane. Once an item is picked up, the user is automatically switched to the loose mode of interaction to allow for free movement while still able to inspect encountered items as their corresponding auditory feedback is displayed upon contact.

Finally, we implemented a synchronisation mechanism to allow the user to switch between the haptic and auditory hierarchical views of the diagrams. The user can locate an item on the hierarchy then issue a command on the keyboard which would cause the PHANTOM arm to move and locate that item on the haptic plane. If the user is holding the stylus, they are then dragged to that location. Similarly, the user can locate an item on the virtual haptic plane then issue a command on the keyboard to locate it on the hierarchy.

3.4. Collaborative Interaction

Simultaneous shared access to a diagram is currently achieved by connecting collaborators' computers through a local network with one of the computers acting as a server. We have incorporated locking mechanisms which prevent collaborators from concurrently editing the same item on a diagram. Besides these locking mechanisms, the tool does not include any built-in mechanisms to regulate collaboration, such as process controls that enforce a specific order or structure of interaction. This was done to allow users to develop their own collaborative process when constructing diagrams – indeed, there is evidence that imposed structure can increase performance but at the expense of hindering the pace of collaboration and decreasing consensus and satisfaction amongst group members [20]. Thus, the cross-modal tool provides collaborators with independent views and unstructured simultaneous interaction with shared diagrams.

4. WORKPLACE STUDIES

We are conducting an ongoing study of cross-modal collaboration between visually impaired and sighted coworkers. The aim is to explore the nature of cross-modal collaboration in the workplace and assess how well the tool we designed supports it in real world scenarios. So far, we have deployed the tool to support the work of three professional pairs; these were employees in the head office of a London-based Children and Families Department in local government, an international charity, and a private business.

4.1. Approach & Setup

We first asked pairs to provide us with samples of the type of diagrams that they encounter in their daily jobs. We then created appropriate templates to accommodate these diagrams on the cross-modal tool. Because we wanted to observe the use of the tool in real world scenarios, involving diagrams of real world complexity, we did not control the type of tasks that the pairs performed nor the way in which they went about performing them. Rather, we deployed the tool in their workplaces and observed their collaborations as they naturally unfolded over a working session. Study sessions lasted for up to two hours, where we introduced the pairs

to the features and functionalities of the tool in the first half, then observed them as they used it to access and edit diagrams in the second half. Visually-impaired participants used the audio-haptic views of the diagrams, where audio was displayed through speakers so that their colleagues could hear what they were doing, while the sighted participants used the graphical view of the tool. In all three cases, the pairs sat in a way that prevented the sighted participants from seeing the screen of their colleagues (see Figure 5), and, naturally, the visually impaired participants did not have access to the graphical view of their partners. We video recorded all sessions and conducted informal interviews with the pairs at the end of the working sessions [3].

Figure 5: An example of the setup used in the workplace.

4.2. Collaborative Scenarios

We observed two types of collaborative scenarios. The first pair, a manager and their assistant, accessed and edited organisation charts to reflect recent changes in managerial structures. The second and third pairs, a manager and an employee assistant and two business partners inspected and edited transportation maps in order to organise a trip. All pairs were able to complete the tasks that they chose to undertake using the cross-modal tool.

Our initial observations showed that collaborations typically evolved over three distinct phases with differing dynamics of interaction. A first phase is characterised as being driven by the visually-impaired user and includes exploring the diagram, editing its content and altering its spatial arrangements. The sighted coworker in this instance typically engages in discussions about the diagram and providing general guidance about where things are located and how to get to them. In a second phase of the collaborations, the visually-impaired user continues to drive the interaction with active input from the sighted user who engages in refining the content and spatial arrangements produced by their coworker. In a third phase, both users engage in manipulating the diagram, working independently on different parts of its content while continuing to discuss the task and updating each other about their progress. These dynamics did not necessarily occur in a particular order. For instance, it is likely that the first phase results from the visually-impaired user's desire to establish orientation within the interactive space at the onset of the collaboration, which might be unnecessary for the sighted user, but such reorientation might occur again after a diagram's content has been extensively altered.

[3] Example videos will be uploaded with the paper and/or shown during the conference presentation.

5. DESIGN RECOMMENDATIONS FOR CROSS-MODAL COLLABORATION

Due to the nature of the study – a small number of participants and uncontrolled workplace environments – we opted for conducting a qualitative analysis of the recorded interactions rather than attempt to capture quantitative aspects of the collaborations. We also focus on aspects of the cross-modal collaborative interaction rather than on the multimodal representation of diagrams. In the following, we present a series of excerpts from the video transcripts [4] to highlight the impact of using audio-haptic displays within the context of cross-modal collaboration and use these example to outline a set of preliminary design recommendations. All videos were transcribed by the first author.

5.1. Extract 1: Exploring and Discussing Diagram Content

In the excerpt shown in Table 1, the pair are editing an itinerary on a transportation map. The excerpt starts off with the visually-impaired user (VI) locating and deleting a node from the diagram while the sighted user (S) edits the label of another node. As soon as the node is deleted, S interrupts VI to inform them about the visible changes that resulted from their action: *"you didn't just delete the node[..]"*. Here the VI user was not aware that deleting a node caused the automatic deletion of the links that were coming in and out of it. The VI user responds with an exclamatory *"yeah?"* while manipulating the haptic device in an attempt to explore the parts of the diagram where the declared changes are said to have occurred. Meanwhile S continues to reason about the outcome of their partner's action: *"we can recreate the .. part of it needed to be deleted anyway"* while the VI user switches to the audio view to check the diagram, correctly deducing the status of its nodes: *"so it only deleted one node.."*.

What we wish to highlight with this excerpts is the way in which the auditory and haptic views were used in the exchange that occurred between the two colleagues. The VI user was able to seamlessly integrate the discussion about the diagram with their partner with the inspection and exploration of the its content. Here, the cross-modal tool formed and effective part of the collaborative exchange; that is, just as S was able to glance at the diagram while discussing and reasoning about its content, so was the VI able to access and explore the diagram while actively partaking in the discussion.

> *Recommendation 1 – Provide explicit representation of the effects produced by a given action to its original author. While the sighted user was able to detect the results of an action as they occurred on the screen, this information was completely oblivious to the original author. It is therefore recommended to explicitly convey the consequences of an action to its original author in the non-visual view. This could also be conveyed in the form of a warning before finalising the execution of an action.*

5.2. Extract 2: Providing Directional Guidance

There were instances in the collaborations where the sighted user provided directional guidance to their partner while they were executing a given editing action. An example of this is shown in the

[4] Since the constructed diagrams were the property of the organisations that we worked with, we deliberately edited out some content and/or concealed it on the transcripts due to the sensitive nature of the information they contain.

Table 1: Extract 1: Smooth embedding of interaction with device and discussion about content.

visually-impaired user	VI actions/audio output	Sighted user	S actions
	<locates node>		
	<deletes node>		<edits node label>
OK, so now I need to			
	<moves the omni>	hold on a second	
what?			
		you didn't just delete the node	
yeah?	<moves the omni>	but also every line that was coming in and out of it	
	<moves the omni>	we can recreate the ...	
	<moves the omni>	part of it needed to be deleted anyway	
		but one didn't	
but that segment had to be removed didn't it?			
let me just .. can i just look for a sec	<explores audio view>		
so it only deleted one node..			
	<explores audio view >	yeah, but every single line ..	

Table 2: Extract 2: Directional guidance.

visually-impaired user	VI actions/audio output	Sighted user	S actions
	<moves the omni to locate a node W>		
	<encounters a node X>		
I've got X			
		then go diagonal left	
	<moves the omni to the left >	up left	
doesn't let me go left	<moves the omni to the left >		
it's literally stopping me from going left	<moves the omni>		
diagonally up or down?	<moves the omni>	up	
from Y or from X?	<moves the omni upwards >		
	<moves the omni>	from X	
	<moves omni to relocate X>		
	<system speaks: "Z">		
		that's the right link, follow Z	
yeah I'm on ..	<follows Z >		
	<locates node W >		

Table 3: Extract 3: Smooth transition between actions.

visually-impaired user	VI actions/audio output	Sighted user	S actions
	<explores the auditory hierarchy>		
	<locates node X and selects it>		
	<explores the auditory hierarchy>		
	<locates node Y and selects it>		
	<creates a link between X and Y>		
	<System confirms the creation of a new link>	alright so I'm gonna move that now	
yup			<selects node X and drags it>

Table 4: Extract 4: Executing a spatial task.

visually-impaired user	VI actions/audio output	Sighted user	S actions
OK, shall we try the others	<moves the omni towards a node>	yup	
	<locates a node X>		
yes, X	<picks up the node>		
got ya	<drags X downwards>		
I'm gonna put it down here somewhere	<drags X downwards>		
What do you recon?		I can't see where you're pointing, drop it first	
I'm gonna put it here	<drops X>		
What do you think?			
		that is again on the same level as the Y	

Table 5: Extract 5: Shared locus.

VI actions/audio output	Sighted user's actions
<edits the label of node X>	<Hovers mouse over node X>
<types new label for X>	<drags X to a new location>
<explores X on the auditory hierarchy>	
<explores X the auditory hierarchy>	<drags X to another location>
<synchronise the audio and haptic views to the location of X>	

Table 6: Extract 6: Exchanging updates.

visually-impaired user	VI actions/audio output	Sighted user	S actions
	<explores the auditory hierarchy>		<edits node Y's parameter>
	<creates a new node X>		<edits node Y's parameter>
	<explores the auditory hierarchy>		<edits node Y's parameter>
	<selects node X on the hierarchy>	so I'm going though Y and Z just adding their details	<edits node Z's parameter>
yeah			
OK	<explores the auditory hierarchy>		<edits node Z's parameter>
I've created the two ...			

excerpt in Table 2. Here, the pair are editing an organisation chart and the visually-impaired user attempts to locate a node on the diagram using the haptic device. The excerpt begins with the VI user moving the device to locate the node in question, encountering an unexpected node X and announcing: *"I got X"*. The sighted user then uses this information to provide their colleague with relevant directions: *"then go diagonal left"*. The VI user attempts to follow their colleague's guidance but, failing to go in the specified direction, seeks more clarification: *"diagonally up or down?"*, *"from Y or from X?"*. Moving around the haptic plan, the VI user encounters another item on the diagram; a link labelled Z. The sighted user picks up on the audio triggered by their partner to tailor the guidance they provide them with: *"that's the right link, follow Z"*. This tailored guidance helps the VI user to locate the node in question.

The fact that the audio output was shared amongst the pair helped the sighted user to engage with their partner's activity. The overlap in presentation modalities in this case created more opportunities for interaction. Information displayed in audio allowed the sighted user to keep track of their partner's progress and, by referring to the graphical view, they were able to map such information and tailor their own discourse to match such progress.

5.3. Extract 3: Transitions Between Collaborative Tasks

The next excerpt, shown in Table 3, shows an example where collaborators executed two dependent actions sequentially. The VI user's task was to create a link between two nodes on the diagram. To achieve this, the VI user first locates the two nodes in question, selects them, then issues a command to create a connection between them. The sighted user's task was to arrange the spatial position of the newly created connection. What is noticeable in this excerpt is that the sighted user was able to determine the exact point in the execution where they were required to take action without being explicitly prompted by their partner: *"alright so I'm gonna move that now"*. Here again, having access to their partner's audio output allowed the sighted user to keep track of their partner's progress resulting in a seemingly effortless transition between the two dependent actions. Thus, allowing an overlap of presentation modalities helps users to structure sequentially dependent actions.

> *Recommendation 2 – Allow an overlap of presentation modalities to increase opportunities for users to engage with each other's actions during the collaboration.*

5.4. Extract 4: Executing a Spatial Task

A major advantage of using a spatial model of representation to support non-visual interaction with diagrams is the ability to execute spatial tasks. The visually-impaired users where able to not only add or remove content from the diagram but also engage with their sighted colleagues to alter content's locations on the diagrams. The excerpt in Table 4 shows an example of this. Here, the VI user uses the omni device to locate a node on the diagram, picks up, drags it across the virtual plane and drops it in a new location. Notice how the VI user engages their sighted partner at each step in the execution of this spatial task by supplying cues about what they are doing: *"yes, X, got ya"*, *"I'm gonna put it down here somewhere, what do you reckon?"*. There is therefore a clear attempt by the VI user to use the spatial layout of the diagram as a common reference when negotiation execution steps with their partner. This was indeed a novelty that was well commended by all participants in our study. The sighted user in the excerpt, however, highlights and important point that contributed to his inability to fully engage with their partner to use this common frame of reference: *"I can't see where you're pointing, drop it first"*. Once the VI user drops the node in the new location it appears on the screen of the sighted user, who could then supply the relevant confirmations to their partner: *"that is again on the same level as the Y"*. Because the tool did not provide the users with any explicit representation of their partner's actions – besides final outcomes – it was hard for them to fully engage with each other during execution. In the case of the excerpt on Table 4, the users compensate for this by supplying a continuous stream of updates of what they are about to do.

> *Recommendation 3 – Provide a continuous representation of partner's actions on the independent view of each user in order to increase their awareness of each other's contributions to the shared space and hence improve the effectiveness of their collaborative exchange.*

5.5. Extract 5: Shared Locus

The excerpt shown in Table 5 does not involve any conversational exchange. However, the pair's interaction with their independent views of the shared diagrams reveals another way in which the two representations were used as a shared locus. In this excerpt, the VI user has created a new node and is in the process of editing its label. Meanwhile, the sighted user moves his mouse and hovers over the node that is currently being edited by their partner then drags it to a new location. The interaction in this excerpt enforces recommendation 2. That is, allowing an overlap of presentation between the visual and audio-haptic display modalities allowed the sighted user to identify the part of the diagram being edited by their partner, to follow the editing process, and to seamlessly introduce their own changes to it (in terms of adjusting the location of the node). The VI user in turn, once finished with editing the label of the node, seamlessly synchronises their auditory and haptic views to explore the new location of the node as introduced by their partner. All of this is done smoothly without any need for verbal coordination.

5.6. Extract 6: Exchanging Updates

The final excerpt in Table 6 shows a different style of collaborative interaction. Instead of waiting for partners to finish executing an action before proceeding with an another, the pair in this excerpt are working in parallel on two independent actions. The VI user in this case is adding new nodes to the diagram and exploring its content using the auditory hierarchical view, while the sighted user is editing nodes parameters. The pair are working in parallel and updating each other about the editing actions that they are currently executing: *"I'm going through Y and Z just adding their details"*, *"I've created the two.."*. Each user is therefore engaged with their own task, and unless an update is supplied, the participants remain unaware of each others progress. Supplying awareness information while both users are jointly engaged with one task is different from supplying it when each one of them is engaged with an independent task. The former, as exemplified in Table 4 was in the form of updates about what the user intends to do, whereas in this excerpt it is in a form of what is currently occurring or what has taking place.

> *Recommendation 4 – While providing a continuous representation of partner's actions, as outline in Recommendation 3 above, care must be taking to choose the most relevant type of awareness information to provide. This changes in accordance with whether the collaborators are executing independent actions in parallel, or engaged in the same dependent tasks in sequence.*

6. CONCLUSION

We presented the design of a collaborative cross-modal tool for editing diagrams which we used to explore the nature of cross-modal collaboration between visually impaired and sighted users in the workplace. An ongoing study that we are conducting in the wild with real world collaborative scenarios allowed us to identify a number of issues related to the impact of cross-modal technology on collaborative work, including coherence of representation, collaborative strategies and support for awareness across modalities. We used our observations to outline an initial set of preliminary design recommendations aimed at guiding and improving the design of support for cross-modal collaboration.

7. REFERENCES

[1] J. Driver and C. Spence, "Attention and the crossmodal construction of space," *Trends in Cognitive Sciences*, vol. 2, no. 7, pp. 254 – 262, 1998.

[2] F. Winberg, "Supporting cross-modal collaboration: Adding a social dimension to accessibility," *Haptic and Audio Interaction Design*, pp. 102–110, 2006.

[3] O. Metatla, N. Bryan-Kinns, T. Stockman, and F. Martin, "Designing for collaborative cross-modal interaction," in *Proc. of Digital Engagement '11: RCUK Digital Economy Community*, 2011.

[4] RNIB, "Looking forward to 2014 rnibs strategy to end the isolation of sight loss," 2009.

[5] G. Kramer, *Auditory Display: Sonification, Audification and Auditory Interfaces*. Reading, MA, USA: Addison-Wesley Publishing Cmpany, 1994.

[6] D. L. Mansur, M. M. Blattner, and K. I. Joy, "Sound graphs: A numerical data analysis method for the blind," *Journal of Medical Systems*, vol. 9, no. 3, pp. 163–174, 1985.

[7] E. D. Mynatt and G. Weber, "Nonvisual presentation of graphical user interfaces: contrasting two approaches," in *Procs. of the SIGCHI'94*, 1994, pp. 166–172.

[8] V. Best, A. Van Schaik, and S. Carlile, "Two-point discrimination in auditory displays," in *Proc. of the 9th Inter. Conf. on Auditory Display*, E. Brazil and B. Shinn-Cunningham, Eds. Boston University Publications Production Department, 2003, pp. 17–20.

[9] G. Weber, "Adapting direct manipulation for blind users," in *CHI '93: INTERACT '93 and CHI '93 conference companion on Human factors in computing systems*, 1993, pp. 21–22.

[10] F. Avanzini and P. Crosato, "Haptic-auditory rendering and perception of contact stiffness," in *Haptic and Audio Interaction Design*, vol. 4129/2006, 2006, pp. 24–35.

[11] W. Yu, K. Kangas, and S. A. Brewster, "Web-based haptic applications for blind people to create virtual graphs," in *Haptic Interfaces for Virtual Environment and Teleoperator Systems. HAPTICS 2003.*, March 2003, pp. 318–325.

[12] G. Leplatre and S. Brewster, "Designing non-speech sounds to support navigation in mobile phone menus," in *Proc. of the 6th Inter. Conf. on Auditory Display*.

[13] E. D. Mynatt and K. W. Edwards, "The mercator environment: A nonvisual interface to x windows and unix workstations, Tech. Rep. GVU Tech Report GIT-GVU-92-05, 1992.

[14] H. Petrie, C. Schlieder, P. Blenkhorn, G. Evans, A. King, A.-M. OâNeill, G. Ioannidis, B. Gallagher, D. Crombie, R. Mager, and M. Alafaci, "Tedub: A system for presenting and exploring technical drawings for blind people," *Computers Helping People with Special Needs*, pp. 47–67, 2002.

[15] O. Metatla, N. Bryan-Kinns, and T. Stockman, "Constructing relational diagrams in audio: the multiple perspective hierarchical approach," in *Proc. of the 10th inter. ACM SIGACCESS conference on Computers and accessibility*, 2008, pp. 97–104.

[16] A. Brown, S. Pettifer, and R. Stevens, "Evaluation of a non-visual molecule browser," in *Proc. of the 6th inter. ACM SIGACCESS conference on Computers and accessibility*, 2004, pp. 40–47.

[17] F. Winberg and J. Bowers, "Assembling the senses: towards the design of cooperative interfaces for visually impaired users," in *Proc. of the ACM conference on CSCW*, 2004, pp. 332–341.

[18] D. McGookin and S. A. Brewster, "An initial investigation into non-visual computer supported collaboration," in *CHI '07 extended abstracts on Human factors in computing systems*, 2007, pp. 2573–2578.

[19] O. Metatla, N. Bryan-Kinns, and T. Stockman, "Interactive hierarchy-based auditory displays for accessing and manipulating relational diagrams," *Journal on Multimodal User Interfaces*, 2011.

[20] J. S. Olson, G. M. Olson, M. Storrøsten, and M. Carter, "Groupwork close up: a comparison of the group design process with and without a simple group editor," *ACM Trans. Inf. Syst.*, vol. 11, no. 4, pp. 321–348, 1993.

EVALUATING LISTENERS' ATTENTION TO AND COMPREHENSION OF SERIALY INTERLEAVED, RATE-ACCELERATED SPEECH

Derek Brock and Brian McClimens

U.S. Naval Research Laboratory,
4555 Overlook Ave., S.W.,
Washington, DC 20375 USA
derek.brock@nrl.navy.mil

S. Camille Peres

University of Houston-Clear Lake,
2700 Bay Area Blvd.
Houston, TX 77058 USA
peressc@uhcl.edu

ABSTRACT

In Navy command operations, individual watchstanders must often concurrently monitor two or more channels of spoken communications at a time, which in turn can undermine information awareness and decision performance. Recent basic work on this operational challenge has shown that a virtual auditory display solution, in which competing messages are presented one at a time at faster rates of speech, can achieve large and significant improvements on diminished measures of listening performance observed in concurrent monitoring at normal speaking rates with equivalent materials. In the third of a series of experiments developed to address performance questions the parameters of this framework raise for listeners, dependent measures of attention and comprehension were compared in a two factor design that manipulated how serial turns among four talkers were organized and their rate of speech. Although both factors impacted performance, the resulting measures remained substantially higher than corresponding measures of performance with concurrent talkers in an earlier study.

1. INTRODUCTION

In Navy command operations, individual watchstanders must often interact with and monitor two or more concurrent channels of spoken radio communications, and this, coupled with the demands of visual tasks, can easily impact information awareness and decision performance [1]. Even so, efforts to increase productivity and streamline operational requirements, have recently raised the possibility of giving watchstanders a range of new display technologies and enlarging their responsibilities to as many as four active communications circuits. A 2001 operational study with a diverse group of experienced watchstanders, however, found that overall message comprehension and awareness of time-critical events fell significantly in a realistic tactical scenario when communications monitoring involving only three channels of competing speech was tasked [2]. This outcome and other findings in the same study suggest that the challenge of attending to multiple streams of concurrent aural information can quickly become overwhelming in high-paced operations.

Monitoring voice communications serially (one at a time) could reduce the considerable requirements of the watchstander's listening task, but would almost certainly result in cumulative and, in some cases, unacceptable presentation delays during periods of high volumes of message traffic. Digitally buffered and recorded speech, however, can be artificially sped up with signal processing techniques that allow the essential timbral features needed for intelligibility and other expressive and informational factors to remain intact. Synthesizing a faster version of what is said on a given radio channel naturally requires a processing delay before it can be aired for the listener—minimally, the time required to receive the original transmission plus a marginal amount of additional processing time. But since competing messages can be processed in parallel, speech rate acceleration techniques can be used to limit the accumulating cost of serial presentation delays and, therefore, provide an opportunity to study serial monitoring as an effective alternative to current communications monitoring practices.

A straightforward model of just under three minutes of activity on four concurrent channels, for instance, would take approximately five minutes to listen to serially, assuming a relatively busy, mean use rate of 40% on each channel and a nearly continuous overlap of two or more messages (see Figure 1a and b). Just doubling the speed of all but the initial message, however, (assuming the first message would be monitored in real-time while competing messages are concurrently buffered and accelerated in parallel) substantially reduces the extent of accumulating delays. Under the acceleration scheme shown here, serialization never adds more than half of the running time required to monitor all four channels concurrently, and

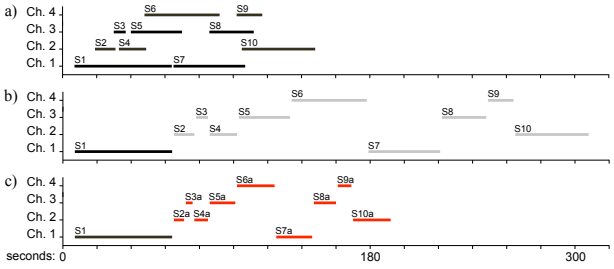


Figure 1: a) Randomized 3-minute model of ten spoken radio transmissions (numbered from S1 in the order they are received) on four concurrent channels. b) Time required to buffer and display the same ten signals serially, in the same order and at the same speaking rate as received (i.e., with no acceleration). c) Time required to buffer, process, and display the same signals at an accelerated speaking rate of 100% (i.e., twice as fast as the original speaking rate). Unserialized messages in the figure (messages presented as they are received) are shown in black. The letter "a" indicates accelerated signals, as in "S2a."

This work was supported by the Office of Naval Research under work request N0001410WX20448.

total listening time is just over three minutes (Figure 1c). More efficient accelerated monitoring and message organization tactics are also possible, but might only rarely be needed due to routine lulls in most real-world patterns of communications traffic.

1.1. Performance concerns for listeners

Although the idea of synthetically accelerating concurrently received messages for subsequent display makes serial monitoring an operationally feasible concept—at least in the sense that it allows serial monitoring to be carried out in nearly the same amount of overall time concurrent monitoring requires—it also raises specific human performance questions for listeners. The most important practical concerns are: a) the performance strengths and weaknesses of human auditory attention; b) performance differences associated with listening to different rates of accelerated speech; and c) the impact of having to shift between communications contexts in an "interleaved" manner, as depicted in Figure 1b and c, or as dictated by some other prioritization scheme.

1.1.1. Auditory attention

Intuitively, listening for content from two or more talkers is harder to do when the parties speak at the same time, as opposed to when they speak individually. Listening to competing talkers requires what is called "divided attention." Both Broadbent [3], and, more recently, Shinn-Cunningham [4], attribute the difficulty of divided listening to an essential limitation of the human auditory attention resource. When divided attention is required, despite anecdotal claims to the contrary, it appears listeners are not really able to focus on two or more auditory streams simultaneously. Instead, while they may be aware of multiple sources and alert to salient features of those sounds, they can only give their attention, selectively, to one coherent stream at a time, and consequently must resort to ad hoc, though possibly practiced, listening strategies that entail rapidly switching their focus back and forth between competing threads of information. What makes giving divided attention to competing auditory streams more difficult than giving sustained attention to one at a time is the mental effort that switching between aural information contexts requires.

As part of a series of experiments that includes the study reported here, Brock et al. [5] examined the question of divided and undivided listening in a quasi-applied context. Working with a corpus of spoken commentaries on everyday topics, inferential measures of auditory attention and comprehension were used to compare listening performance in four manipulations involving either concurrent or serial talkers. The manipulations with concurrent talkers (two talkers in one condition and four in the other, and both at normal speaking rates) reflected current and proposed Navy communications monitoring practices. The serial talker manipulations (one at normal speaking rates, the other at an accelerated rate of 75%, and both with four talkers) allowed serial monitoring to be compared directly with concurrent listening and provided an initial look at the impact of accelerated speech on serial listening performance. The resulting measures of attention and comprehension proved to be highly correlated with each other, and all pairwise comparisons between the manipulations were significant. Listening performance was respectively poor and poorest in the two and four concurrent talker conditions, and better and best in the accelerated and normal serial conditions. The outcome was thus consistent with the current understanding of auditory attention and demonstrated a clear performance advantage for serial monitoring over current practice, even with faster speech.

1.1.2. Rate-accelerated speech

Techniques for synthetically compressing (and, therefore, accelerating) the nominal speaking rate of normal, recorded speech—without altering its pitch—were first studied in the early 1950s. Research by Miller and Licklider [6] showing that brief segments of continuous speech could be either systematically blanked out ("interrupted") or masked with only modest impacts on perceived intelligibility led to the idea of splicing together what remained to reduce listening time [7]. Eventually, as interest in accelerated speech grew, and access to digital signal processing technology became widespread, more sophisticated speech-rate modification techniques were developed that are capable of preserving most, if not all, of the vocal features involved in clear enunciation at rates of compression that exceed 200% (see [8] for an outline of research up to the beginning of the 1990s). The technique used in the work reported here is a computationally efficient method for modulating the time scale of speech known as "pitch synchronous segmentation" (PSS) that was developed by the Navy in 1994 [9]. Human performance and perceptual studies associated with rate accelerated speech have focused primarily on the intelligibility of individual words and the practical limits of acceleration, as well as the impacts of acceleration and prosodic modifications (particularly, the removal of pauses) on the more practical question of comprehension performance. Additional work has also explored the impacts of training and practice and, more recently, performance differences associated with aging (see, e.g., [10]).

Since varied pacing might be needed to accommodate changing amounts of message traffic in a serial communications monitoring scheme, two experiments ([1] and [11])—one planned as a follow on for [5] and another developed by Wasylyshyn—were recently conducted to examine listening performance with different rates of accelerated speech using the PSS technique [9]. Although different materials and exposure regimes were used in each protocol, the outcomes of both studies are in general agreement with the findings of earlier research on this question using other speech rate compression methods. Brock et al. [1] found comprehension of compressed speech up to a 100% increase in ordinary speaking rates (i.e., twice as fast) to be essentially equivalent to listening to normal speech. Similar equivalence was observed in Wasylyshyn's study [11] up to a rate of 80%, or 1.8 times as fast as normal speech. Above these levels, as in other research, performance was found to slowly but significantly decline in a relatively steady manner as the degree of acceleration grows. In both studies, however, even at the highest levels of accelerated speech rates (175% in [1] and 140% in [11]), mean comprehension was much better than, or as good as, the listening conditions involving two and four concurrent talkers in [5]. The consistency of these performance

outcomes with the findings of others suggests that the ability of listeners to follow and make verifiable sense of synthetically accelerated speech at speeds up to and beyond a 100% increase in normal speaking rates is a readily acquired skill.

1.1.3. Listening to serially interleaved communications

Questions concerned with interleaving, specifically, shifting back and forth between communication contexts, are motivated by the insight that competing communications are just that. If one message is more timely or important than another, the listener will want to give its presentation priority, even if this means withdrawing attention from or suspending the less urgent of the two and returning to it later. Suspension would be the case in a serial monitoring scheme, and the issue then becomes, what is the likely impact of system-imposed interruptions on listening performance when messages are subsequently resumed. Even more to the point in a communications setting is the fact that what is said on most radio nets is not just one individual talking, but discourse among multiple talkers. Thus, upon resuming a suspended channel, the listener not only faces the problem of attending while reengaging with the channel's operational context and recalling its state, but also of recognizing who the talker is and/or what the talker's role in the current communications context is. Mastering these additional aspects of the serial listening task may well be made more difficult by accelerating what is displayed for the listener, even if the increase falls within the range of equivalent-to-normal comprehension performance.

However, other factors may measurably impact listening performance, too. The most important concerns are: message complexity; where suspensions occur within a message stream; and whether or not the pace of display provides opportunities to reflect on or rehearse a suspended context before it is resumed. For instance, in addition to difficulties that ordinarily arise for listeners when speech materials in any format are syntactically complex (e.g., [10]), listening performance is known to be hurt when unexpected pauses occur in sentences, as opposed to at grammatical clause boundaries [12]. From this, it follows that listeners are likely to find arbitrary suspension points more difficult to work with than suspensions that occur at the end of clauses or on sentence boundaries, or perhaps breaks that occur between different talkers on a given channel.

As for the pace of display, listening that involves interrupting one informational context and attending to another can be likened to a sequential multitasking paradigm [13]. Current cognitive theories of multitasking model the ability to juggle more than one task at a time as interacting goal hierarchies [14] and, more recently, as separate "threads" of goal directed activity [15]. For comprehension tasks, people often need to maintain an informational context or "problem state"—a small amount of applicable knowledge, and/or intermediate results, that is temporarily buffered for working access. Recent work by Borst et al. [16] has concluded that the cognitive resource for this intermediate store can only be used by one task thread at a time. Thus, part of the difficulty of managing even two ongoing comprehension tasks at once, whether they are perceptually concurrent or sequentially paced, is explained by the mental effort that is needed to repeatedly reinstate their respective contexts. The time this requires can become an issue, too, if the wait before the next episode of attention to a task becomes too long. In a serial monitoring scheme, a progression of different channels may intervene before a given interrupted channel is resumed, depending on how the incoming spoken information is prioritized and segmented. As the pace of imposed switching between suspended contexts slows, listeners will have increasing difficulty recalling each channel's respective problem state [17]. Empirical studies and related modeling work by Trafton et al. [18] and Altmann and Trafton [19] have shown that to counteract this quantifiable tendency to forget, listeners need to rehearse an interrupted context—ideally, at the point when the interruption occurs. Consequently, if in addition to relatively slow pacing, switches between channels are effectively immediate (with no gap to briefly think about what was just interrupted), listening performance can be impacted in two ways. Either, listeners will try to rehearse the previous context anyway, and initial attention to the new context will be impaired, or listeners will fail to think about the previous context and have greater difficulty recalling it later, which will also impair initial attention to the new context.

1.2. Listening study

The listening study reported here—an initial 2x2 comparison of interleaved and non-interleaved listening with normal and rate-accelerated speech—is the third in a series of experiments that includes the work presented in [5] and [1]. For consistency with the previous studies, the speech materials used for auditory display in the present experiment were again developed from a public radio archive of spoken essays by four professional commentators. (An essay from an additional commentator was also used for training purposes—see Section 2.1.3 below). This category of talk sidesteps potential confounds and has specific advantages for the population of non-specialist listeners recruited to participate in the study. In particular, each commentary is presented by a single talker and, so, avoids contextual confusions that could arise from the presence of more than one voice on the same channel. Each commentary also covers a single everyday topic in ordinary conversational language that is easy to follow and quickly establishes an easily recognized contextual theme for the channel it corresponds to during its presentation.

A serially interleaved communications display in an operational setting would probably exhibit some of the characteristics depicted in Fig. 1c, notably, a mix of communications sounded at normal and accelerated rates and a mixed range of message lengths. The present study's chief aim, however, was to examine the impact of interleaving itself on listening performance with normal and faster speech, as opposed to other issues interleaved listening designs may raise. Consequently, the main questions addressed here are: a) Is the problem of having to follow and understand four different spoken information contexts harder to do when, instead of being allowed to listen to each talker's full presentation, one at a time, what each talker has to say is broken into an ordered series of utterances that are displayed as a randomized sequence of turns among the talkers? And b) how does making the speech materials in these contrasting conditions much faster affect the ability of listeners to follow and understand all of what each talker has to say?

Because serially interleaved listening can be characterized

as an example of sequential multitasking, the performance concerns raised in Section 1.1.3 related to the interruption of contexts are addressed in the experimental design by the insertion of a brief gap after each talker's turn in the manipulations that involve interleaved listening. The intent in doing this, though, was not to measure the impact of remedial measures for interruptions, but rather to organize the design of the interleaved listening task in a theoretically principled way. To ensure talkers had equal priority throughout, each commentary was edited to approximately the same length and segmented into a congruent (equally numbered) sequence of utterances or "turns." Four commentaries (one per talker) were presented in each of the listening exercises, and in those with interleaved utterances, the order of sequential turns among the talkers was randomized for each listener (see Section 2.1.3 below for additional details). As a result, the wait between a given talker's completed turn and that talker's next turn in the interleaved listening exercises might be short or long, but, on average, entailed the span of time defined by the first inserted gap, plus three turns from the other talkers, plus the gaps inserted after each of these intervening turns. Each gap thus provided a moment to think about the completed turn's context, but for consistency with the non-interleaved portion of the study, no constraint was placed on how listeners were expected to manage their time during any of the listening exercises.

Other aspects of the experimental task design that were similarly informed by current theory are the use of separate virtual locations for each talker in the auditory display and the manner in which commentaries were divided into turns. Giving the apparent source of each talker's voice its own virtual location, and keeping this constant throughout the study, provided two, closely related theoretical benefits for listeners. First, it capitalized on the spatial skills listeners routinely use to discriminate between sources of auditory information in selective attention (cf. [20]). And second, it provided an external set of talker-specific, contextual cues in the aural information environment. That is, listeners could use each talker's virtual location as an aural reminder for returning to that talker's corresponding problem space during the serially interleaved listening exercises. (Listeners were also able to exploit an external set of visual cues in these exercises; see Section 2.1.1 and 2.1.2 below.) As for turns, speech on competing radio channels could no doubt be broken into separate utterances in several different ways for interleaved display in an operational serial monitoring scheme. Empirical findings such as those in [12], however, suggest that forming an understanding of interrupted speech is facilitated when interruptions occur on grammatical and/or conceptually complete boundaries, and that listeners perform best when this is the case. Thus, to minimize performance confounds related to encoding, in addition to dividing each commentary into an equal number of turns, all of the partitions were made so utterances were sentences or complete phrases.

2. METHOD

Sixteen participants, two female and fourteen male with a mean age of 29.3 years (s.d. = 10.7), all personnel at NRL, and all claiming to have normal hearing, took part in the experiment, which employed a within-subjects design. The visual part of the study was displayed on an NEC MultiSync LCD 2090UXi flat-panel monitor and the aural component was rendered with VRSonic Vibestation runtime spatial audio software, Sony MDR-600 headphones, and an InterSense InertiaCube3 for head tracking. The main experiment consisted of four listening exercises, which were performed by all participants in counterbalanced order. A brief introductory session before the study explained each of the ways participants were asked to respond and described what they would hear and see in the study. Each condition in the main experiment was preceded by a short training session that resembled the format of the listening exercise that followed. These sessions allowed participants to become familiar with the auditory manipulations and their corresponding listening requirements.

2.1. Apparatus

Listeners were asked to make two types of responses in the experiment—the first while listening and the other performed immediately after. Both of these tasks are largely the same as those used to assess listening performance in [5] and [1].

2.1.1. Response tasks

In the first response task, participants were instructed to mark items in a set of lists that were displayed on the flat-panel monitor during the auditory portions of both the training sessions and the main listening exercises. Each list (as well as its left-to-right position onscreen) corresponded to one of the commentaries being presented in the current segment of the experiment and was composed of an ordered set of noun phrases. There were four lists and four commentaries in each of the main experimental manipulations and two lists and two commentaries in each of the respective training sessions. Each list functioned as a visual contextual cue when its talker's commentary was active. Participants were asked to use a mouse to successively check off exactly worded phrases if they heard them spoken (targets) and to ignore any intervening, though topically similar, phrases they did not hear (foils). Lists in the main listening exercises were each composed of twenty targets and an equal number of foils, with zero to three intervening foils placed at random between targets, and no more than three targets in a row. (Shorter lists were used in the training sessions.) In part, because participants were not made aware of the arrangement of targets and foils, and in part because of the potential to become lost while trying to perform the phrase recognition task (thus, undermining the overriding goal of listening), a portion of the currently active list was highlighted as a pale blue region that functioned as a position marker corresponding roughly to the utterance that was currently being presented in the active commentary (see Figure 2a). To ensure that listeners could not game the task, the highlighted area moved continuously and always encompassed several phases in the active list.

In the second response task, which is derived from a technique for measuring reading comprehension developed in [21], participants were given a series of representative sentences to read and asked to judge whether each contained "old" or "new" information based on the spoken materials they had just listened to. "Old" sentences were of two types: verbatim renderings and synonymous paraphrases of sentences in the commentaries. "New" sentences were similarly of two

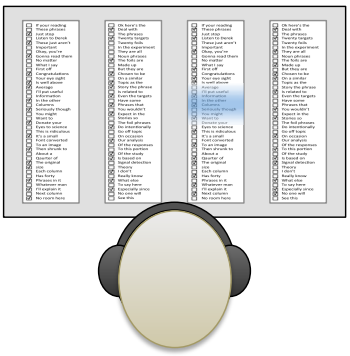
types: "distractors"—sentences stating something that was not implied or said—and commentary sentences with one or two words changed to make the meaning clearly different from what was said. Participants were also given the option to indicate that they did not know whether a sentence they were asked to evaluate was old or new. In the training sessions, participants were given only two sentences per commentary to assess, one old and the other new. Eight sentences per commentary (two of each of the old and new sentence types) were given in the main listening exercises.

In the present study, participants were also asked to indicate how confident they were in their judgments. They did this with an appropriately labeled onscreen widget resembling a slider, with end points labeled "Low" and "High." When participants indicated they could not evaluate a particular sentence, the confidence scale was grayed out.

2.1.2. Auditory display

All of the auditory manipulations were presented in a virtual listening environment organized somewhat similarly to the auditory displays used in [5] and [1]. In this experiment, however, head tracking was also used to implement an augmented auditory reality display, meaning that the apparent referential frame of the virtual aural setting remained the same as that of the actual visual setting, regardless of how participants moved their heads. Each of the normally spoken and rate-accelerated commentaries was binaurally filtered and rendered with headphones using a non-individualized head-related transfer function. To ensure that participants could quickly focus their aural attention on the active commentary (cf. [20], [22]), the apparent locations of the four talkers in each of the main listening exercises were positioned, from left to right on the virtual horizontal plane in front of the listener, at -75°, -25°, 25°, and 75°, with 0° being straight ahead in the visual environment (see Fig. 2b). In the training sessions, only the -25° and 25° positions were used. Each talker's virtual location was maintained across all manipulations and, as was noted above, each of these locations corresponded in a left-to-right manner to the visual location of its matching phrase list in a given exercise on the flat screen monitor.

a)



b)

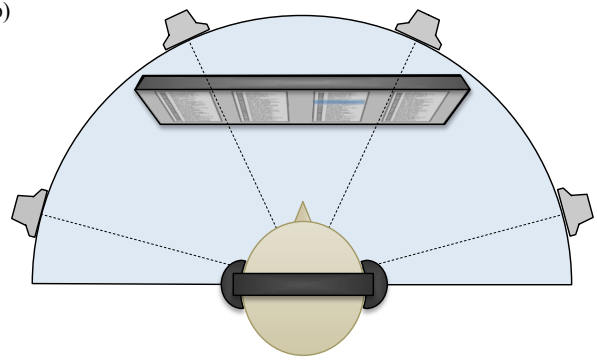


Figure 2: a) Illustration of the visual display showing the four lists of target and foil noun phrases used for the phrase recognition task in the listening exercises. Each list corresponds to a talker, and the pale blue region about midway down the third list indicates that the middle-right talker is speaking. Listeners were asked to mark any noun phrases in each list they heard the corresponding talker say. b) Diagram of the runtime spatial audio environment showing the virtual locations of the four talkers in the listening exercises and their left-to-right correspondence with the onscreen lists used for the phrase recognition task.

2.1.3. Listening materials and experimental manipulations

Each participant in the study listened to a total of 18 spoken essays by two female and three male commentators selected from an internet archive of public radio broadcasts. Both of the women and two of the men were designated as the set of talkers participants would hear in each of the study's main listening exercises. Four pieces from each of these individuals were chosen and edited to remove music and other non-speech sounds. The resulting 16 commentaries ranged from 2 min. 9 sec. to 2 min. 32 sec. in length, with a mean length of 2 min. 19 sec. Listeners heard one commentary per talker in each of four experimental manipulations in the main body of the experiment. In addition to these commentaries, two shorter pieces were also selected and similarly edited for the study's training sessions. Both were spoken by male talkers, of whom, one was the remaining male commentator from above. Participants trained with appropriately manipulated versions of these two commentaries before each of the main listening exercises. These short training sessions allowed participants to become acquainted with the format of each of the auditory display manipulations and practice the listening requirements.

All of the commentaries were further edited into ordered sequences of successive, non-overlapping clips, with each clip corresponding to an utterance. The edits were made so that utterances were either complete sentences or grammatically complete clauses. Additionally, each utterance was edited to start and end with its talker's voice, meaning that any preceding or trailing silence at these specific points was removed. The 16 commentaries used in the main listening exercises were divided in to 15 clips each, with utterances ranging from 4 to 16 sec. and averaging about 9.5 sec. The short commentaries used for the training sessions were also similarly divided into six clips each. Next a version of each clip at double the rate of its original speech was generated with the PSS algorithm [9]. 100% acceleration was chosen for the study because listening performance at progressively faster rates of speech markedly declined above this point in [1].

Each of the four main listening exercises implemented a separate treatment within a two-factor, 2x2, repeated measures design. The first factor, presentation, manipulated the serial organization of talker turns (two levels: **Full** turns, with each

Condition	Description
FN	**Full** turns, **Normal** speech
FA	**Full** turns, **Accelerated** speech (100% faster)
IN	**Interleaved** turns, **Normal** speech
IA	**Interleaved** turns, **Accelerated** speech (100% faster)

Table 1: A summary of the four experimental conditions and their coded designations.

turn being a full commentary vs. **Interleaved** turns, with each turn being an utterance). The second factor manipulated each talker's speaking rate (two levels: **Normal** speech vs. **Accelerated** speech). Table 1 summarizes the manipulations in each of the four conditions and serves as a key for their coded designations in the remainder of the paper. Overall listening performance was predicted to be best in condition **FN**, and progressively worse in conditions **FA**, **IN**, and **IA,** in that order.

The treatments and listening materials were organized in the following way. The 16 commentaries developed for the main listening exercises were divided into four groups of four commentaries with one from each of the four talkers. These four groups were used for the four listening exercises each participant carried out. Participants were assigned to one of four different treatment orders based on a 4x4 latin square, in the order of their enrollment. Further, to ensure that all pairings of treatments and commentaries appeared in the study an equal number of times, each order of treatments was combined with a different ordering of the four commentary groups (also based on a 4x4 latin square).

Silent pauses of pre-defined lengths were inserted between clips at runtime in each of the listening exercises, as well as in the training sessions, to simulate natural pauses talkers frequently add between clauses and sentences in normal speech. The lengths of inserted pauses were proportional to the speed of the speech materials: 400 ms was used for pauses in normal speech and 200 ms for pauses in accelerated speech.

In each of the listening exercises involving full turns (the **FN** and **FA** conditions), commentaries were presented from left to right. Thus, the sequence of clips corresponding to the first talker's full commentary were played in order, with pauses inserted between them, followed by the next talker's full set of clips and inserted pauses, and so on, until all four commentaries had been aired. In contrast, in the listening exercises involving interleaved turns (the **IN** and **IA** conditions) the following algorithm was used to alternate among each of the talkers' commentaries: The first talker was chosen at random, and the first clip from the corresponding commentary was removed from its sequence of utterances, played with a pause inserted at the end, and followed by an additional gap of 300 ms (this is the "brief gap" discussed in Section 1.2 above). This set of actions completed the first "interleaved" turn. Each successive clip was then selected from the commentary with the greatest amount of time remaining and played in the same manner as the first clip. In the event of a tie (e.g., two or more commentaries had an equal amount of time remaining), the next talker was again chosen at random. This procedure continued until all four sequences of utterances were exhausted. The addition of the 300 ms gap after each interleaved clip and its inserted pause made the net pause between interleaved utterances 700 ms in the **IN** condition and 500 ms in the **IA** condition. 300 sec. gaps were not added in the **FN** and **FA** manipulations because full turns allowed listeners to focus on each talker for over two minutes at a time, and all of the commentaries had a clear beginning, middle, and end.

2.2. Dependent Measures

In the series of studies this experiment is part of, the participant's task of listening for information is regarded as having two successive stages of perceptual performance: aural attention and aural comprehension. Neither of these functions is directly observable, so indirect techniques are needed to estimate how well the listener discharges them. As in [5] and [1], phrase recognitions and sentence judgments are used for this purpose.

2.2.1. Attention

The first response task, which required participants to recognize specific noun phrases in the speech materials (see Section 2.1.1), is used here as a measure of attentional performance—specifically, how well listeners were able to attend to and identify what each of the talkers said during the listening exercises. The use of targets and foils in this task allows performance to be scored in two ways—as a proportion of correctly identified targets and rejected foils and, alternatively, as a *d′*. The latter, which is reported here, is a signal detection sensitivity score derived from the respective rates of "hits" (targets correctly identified) and "false alarms" (foils marked as targets). *d′* can be thought of as the distance between the means of the observed distributions of hits and false alarms. Higher values for this measure indicate that listeners marked many of the targets and very few of the foils[1].

2.2.2. Comprehension

Aural comprehension performance is measured here as the combined proportion of sentences participants correctly judged to be consistent or inconsistent (i.e., "old" or "new"; see Section 2.1.1) with the speech materials they had just listened to in each of the experimental manipulations. Because a strong correspondence between respective patterns of attention and comprehension performance was previously observed in this series of experiments (see [5] and [1]), a similar correspondence was expected in the present study. Other measures associated with listeners' sentence judgments are their confidence scores—a self-reported measure of how certain they were about each judgment, ranging from "not at all" to "very" (see Section 2.1.1)—and the number of "I don't know" responses each listener made. Analyses of these data will be reported elsewhere.

3. RESULTS

A two-factor, repeated measures analysis of variance, with two levels for each factor (presentation: Full vs. Interleaved turns;

[1] *d′* was calculated with substitute fractional rates of 1-(1/(2N)) and 1/(2N) for listeners with a perfect hit rate of 1 and/or a false alarm rate of 0, using the number of targets or foils for N.

and speaking rate: Normal vs. Accelerated speech), was performed for each of the dependent measures derived from the response task data. Performance in each of the treatments was largely consistent with the expected pattern of differences.

3.1. Attention

There were significant main effects of speaking rate and presentation on participants' *d'*s, which index aural attention performance (the ability to follow what was said) in terms of how often participants chose targeted noun phrases and incorrectly chose foils as they listened to the commentaries. Specifically, phrase discrimination was hurt by accelerating the rate of speech, regardless of whether talkers took full or interleaved turns (F (1, 15) = 98.15, $p < 0.001$, $\eta^2 = 0.867$). Additionally, performance fell when talkers took interleaved turns, regardless of the rate of speech (F (1, 15) = 4.98, $p = 0.041$, $\eta^2 = 0.249$). There was no interaction between the factors ($p = 0.72$). Figure 3 shows mean d' scores plotted by presentation and speech rate.

3.2. Comprehension

Participants' mean scores in the comprehension response task are given in Figure 4. The proportion of correct sentence judgments participants made after listening to the commentaries in a given exercise dropped significantly when the rate of speech was doubled (F (1, 15) = 37.8, $p < 0.001$, $\eta^2 = 0.716$). As the plots in the figure show, accelerated speech undermined how well participants were able to decide if representative sentences were consistent with their understanding of the speech materials when talkers took full and interleaved turns. In contrast, there was no main effect of presentation—comprehension performance was not significantly impacted when commentaries were displayed as a series of interleaved turns among talkers ($p > 0.10$). Additionally, there was no interaction between factors ($p > 0.10$).

4. DISCUSSION AND CONCLUSION

The first and most pressing question the present study intended to address is the effect of serial interleaving (dividing what multiple talkers have to say at the same time into an alternating sequence of turns) on the ability of listeners to keep track of and

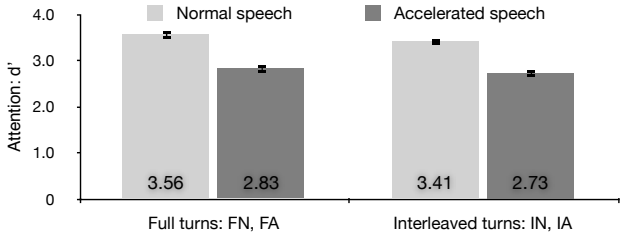


Figure 3: Mean aural attention performance, indexed by the signal detection score *d'*, showing the extent of participants' ability to recognize targeted noun phrases and minimize the selection of foils (phrases not present in the speech materials) while listening in each of the experimental treatments. Higher scores indicate better performance. Error bars show the standard error of the mean (s.e.m.).

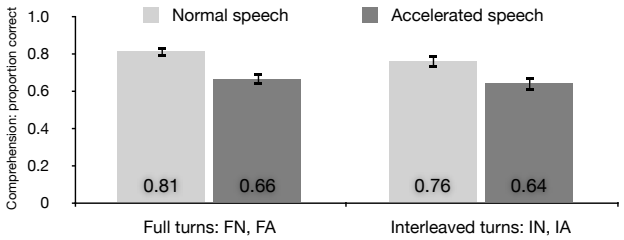


Figure 4: Mean aural comprehension performance as measured by the proportion of representative sentences participants correctly judged as consistent or inconsistent with their understanding of the spoken materials after listening in each of the experimental treatments. Error bars show the s.e.m.

understand the import of each thread of spoken information. The motivations for examining this way of intercepting and organizing competing contexts of speech are the inherent performance costs of attending to them at the same time and, conversely, the likely operational drawbacks of listening to each at length and one-at-a-time.

Serially interleaved listening reconciles the requirements of competing information priorities and alleviates the more difficult work of divided attention by allowing one context to be interrupted by another and resumed later. However, it also poses all of the challenges of sequential multitasking for listeners. Consequently, listening performance in the study's comparison of commentaries spoken in full turns and in interleaved turns was expected to be somewhat worse in the latter two manipulations because of the disruptive effects of repeated interruptions. As it turned out, though, while interleaving did have a significant impact on listeners' aural attention scores, the effect was not large, and, surprisingly, there was no corresponding effect of interleaving on listeners' comprehension performance at all.

Several theoretically motivated elements in the design of the listening task (outlined above primarily in Section 1.2) may have contributed to this outcome, including: the insertion of 300 ms gaps between interleaved turns; the external contextual cues provided by each talker's aural location and corresponding onscreen phrase lists (as well as linguistic cues in these displays); how the commentaries were divided into separate utterances; and the wide spatial separations between talkers in a stable virtual listening environment. If this is the case, it suggests that while serial interleaving necessarily imposes attentional costs on listeners, it can, in fact, be designed and displayed in ways that help to ameliorate the more decisive performance tolls sequential multitasking can potentially levy on tasks, particularly, functional loss of contextual understanding.

Although the second outcome of the study—the significant impact of accelerated speech on both measures of listening performance, regardless of how turns were organized—was not wholly unexpected, it also included an unanticipated development that may be a consequence of workload and how performance was measured. The decision to compare listening to normal and 100% faster speech in the study's design was made on the premise that acceleration rate should be at or just above the range where empirical performance begins to fall (e.g., per [1]). Moreover, because interleaving and accelerated speech were both expected to produce performance effects, an

important aim of the study was to evaluate how profoundly the upper end of effective accelerated speech might hurt serially interleaved listening performance. What was unexpected was that accelerated speech, rather than interleaving, would be responsible for the largest effects in the study (thus the anticipated order of performance declines across manipulations given in Section 2.1.3). A plausible explanation for this result, though, may be tied to differences in the respective ways aural attention and comprehension were measured here and in [1] and [11]. In [1], in particular, the method and specific manipulations were much the same as the **FN** and **FA** treatments above. However, in [1], participants only listened to two commentaries per exercise, which suggests that the use of four talkers per exercise here may have increased the workload associated with the response tasks enough to impair both measures of performance with faster speech at or near previously observed ceilings.

Still, to place the study's key performance result in context, it is worth noting that while the combined impacts of interleaving and accelerated speech respectively reduced attention and comprehension performance in the **IA** manipulation to a mean *d′* of 2.73 and to a mean proportion of correct sentence judgments of 0.64, both of these scores are substantially higher than the corresponding scores for the two and four concurrent talker conditions reported in [5]. In those manipulations, listeners' mean *d′*s were 1.93 and 1.45, respectively, and their mean proportions of correct sentence judgments were respectively 0.47 and 0.25.

Analyses of additional measures collected in the study will be reported at a later date. Future research on the applied use of this framework should begin with issues raised by more operationally realistic speech materials and performance questions raised by its integrated use in a mixed-purpose auditory display.

5. REFERENCES

[1] D. Brock, C. Wasylyshyn, B. McClimens, and D. Perzanowski, "Facilitating the watchstander's voice communications task in future Navy operations," in *Proceedings of the 2011 IEEE Military Communications Conference (MILCOM)*. Baltimore, MD. November 7-10, 2011.

[2] D. Wallace, C. Schlichting, and U. Goff, Report on the Communications Research Initiatives in Support of Integrated Command Environment (ICE) Systems, Naval Surface Warfare Center Dahlgren Division, TR- 02/30, January, 2002.

[3] D. W. Broadbent, *Perception and Communication*, Pergamon Press, New York, NY, USA, 1958.

[4] B. G. Shinn-Cunningham, "Object-based auditory and visual attention," *Trends in Cognitive Sciences*, 12, 182-186, 2008.

[5] D. Brock, B. McClimens, J. G. Trafton, M. McCurry, and D. Perzanowski, "Evaluating listeners' attention to and comprehension of spatialized concurrent and serial talkers at normal and a synthetically faster rate of speech," in *Proceedings of the 14th International Conference on Auditory Display (ICAD)*, Paris, France, June, 2008.

[6] G. A. Miller and J. C. R. Licklider, "The intelligibility of interrupted speech," *J. Acoustical Soc. Am.*, 22(2):167–173, 1950.

[7] W. D. Garvey, "The intelligibility of speeded speech," *J. Exp. Psychol.*, vol. 45, no. 2, pp. 102–108, 1953.

[8] B. Arons, "Techniques, perception, and applications of time-compressed speech," *Proc. 1992 Conf., Am. Voice I/O Soc.*, pp. 169-177, 1992.

[9] G. S. Kang and L. J. Fransen, "Speech Analysis and Synthesis Based on Pitch-Synchronous Segmentation of the Speech Waveform," Naval Research Laboratory, TR-9743, November, 1994.

[10] A. Wingfield, J. E. Peelle, and M. Grossman, "Speech rate and syntactic complexity as multiplicative factors in speech comprehension by young and older adults," *Aging, Neuropsychology, and Cognition, 10,* 310 –322, 2003.

[11] C. Wasylyshyn, B. McClimens, and D. Brock, "Comprehension of speech presented at synthetically accelerated rates: Evaluating training and practice effects," in *Proceedings of the 16th International Conference on Auditory Display (ICAD)*, Washington, DC, June, 2010.

[12] S. S. Reich, "Significance of pauses for speech perception," *J. of Psycholinguistic Res.*, 9(4):379–389, 1980.

[13] D. D. Salvucci, N. A. Taatgen, and J. P. Borst, "Toward a unified theory of the multitasking continuum: From concurrent performance to task switching, interruption, and resumption," in *Human factors in computing systems: CHI 2009 conference proceedings*, New York, NY: ACM Press, 2009, pp. 1819–1828.

[14] D. E. Kieras, D. E., Meyer, J. A. Ballas, and E. J. Lauber, "Modern computational perspectives on executive mental processes and cognitive control: Where to from here?" in S. Monsell & J. Driver (Eds.), *Control of Cognitive Processes*, Cambridge, MA: MIT Press, 2000 pp. 681-712.

[15] D. D. Salvucci and N. A. Taatgen, *The Multitasking Mind*, Oxford University Press, 2011.

[16] J. P. Borst, N. A. Taatgen, and, H. Van Rijn, "The problem state: A cognitive bottleneck in multitasking," *J. Exp. Psychol.: Learning, Memory, & Cognition*. vol. 36, no. 2, pp. 3363-382, 2010.

[17] E. M. Altmann and J. G. Trafton, "Memory for goals: An activation-based model," *Cognitive Science*, 26, 39-83, 2002.

[18] J. G. Trafton, E. M. Altmann, D. P. Brock, and F. E. Mintz, "Preparing to resume an interrupted task: Effects of prospective goal encoding and retrospective rehearsal," *Int. J. of Human-Computer Studies*, 58, 583-603, 2003.

[19] E. M. Altmann and J. G. Trafton, "Timecourse of recovery from task interruption: Data and a model," *Psychonomic Bulletin & Review*, 14, 1079-1084, 2007.

[20] V. Best, F.J. Gallun, A. Ihlefeld, and B.G. Shinn-Cunningham, "The influence of spatial separation on divided listening," *J. Acoust. Soc. Am.*, vol. 120, no. 3, pp. 1506–1516, Sept. 2006.

[21] J.M. Royer, C.N. Hastings, and C. Hook, "A sentence verification technique for measuring reading comprehension," *J. Reading Behavior*, vol. 11, no. 4, pp. 355–363, 1979.

[22] A.W. Mills, "On the minimum audible angle," J. Acoust. Soc. Am., vol. 30, no. 4, pp. 237–246, Apr. 1958.

A PERSPECTIVE ON THE LIMITED POTENTIAL FOR SIMULTANEITY IN AUDITORY DISPLAY

Joachim Gossmann

UC San Diego
Center for Research and Computing in the Arts
9500 Gilman Drive La Jolla, California 92093-0037
jgossmann@ucsd.edu

ABSTRACT

The auditory environment has been described as a *biased competition*: The juxtaposition of an array of pre-formed auditory streams and a process of attentional selection [1, 2]. The orientation of attentional selection toward environmental streams is differentiated towards different *modes* of streaming: Speech, music and sound effects are only three examples in a potentially open polymorphism of *perceptual strategies* through which we access the sounding world.

This differentiable-simultaneous manifold of environmental streams allows perceptual participation only within a certain number of processes at the same time—only one speaking voice, one sense of ”harmony”, a single ”rhythm”, and so forth.

We propose a re-basing of sonification strategies not on the definition of external mechanisms, but on the definition and application of new *modal strategies* that are circumscribed and accessible through *what is not possible to perceive at the same time*.

1. INTRODUCTION: THE DIFFERENTIABLE-SIMULTANEOUS MANIFOLD

The phenomenon of *multiple parallel channels of information* encounters on many structural levels in time-based media artifacts: The distinct intertwined voices of contrapuntal music, the parallel polymorphism of dialog, music and sound effects projected onto the audience from a multichannel loudspeaker array in movie soundtracks, radio drama and news reports that combine location-sound with added voice-over, and the two ears that we both hear with at the same time. We find ourselves addressed by representations and expressions of a multiplicity of simultaneously present streams, objects and events. Auditory media, which unfold exclusively in temporal developments, seem to imply the potential to display a manifold of simultaneous signals and processes to the participant. But before we can approach a structural description of the phenomena of perceptual simultaneity, we should first generate transparency in an area of potential misunderstanding.

1.1. The distinction between audio channels, sensory channels and environmental streams

We can distinguish three structural levels on which we find arrays of parallel streams:

1. the *array of audio channels* that are stored and transmitted by the medium and projected by the loudspeakers or headphone
2. the *sensory array* of the participant
3. the *manifold of environmental streams* that make up the auditory scene the observers and participants find themselves immersed in

Evidently, we find the polyphonies we experience in the audio content itself (layer 3 in this model) encoded and transmitted through layers 1 and 2. However, each of these connected layers is characterized by the a potential for structural independence. Especially the relationship between a loudspeaker signal and an environmental stream is a source of potential confusion. We usually do not encounter the voices that make up a musical polyphony projected from distinct physical sources, channels or spatial locations—a string quartet represented by four discrete loudspeakers for example. Instead, the count of transmitted and projected media channels tends to conform to the properties of the sensory array of the participant—stereo loudspeakers, headphones, (video screens in the audio-visual case, sometimes with two simultaneous images, one for each eye). But we are increasingly confronted with cases in which the count of discrete audio channels that are projected from loudspeakers is greater than the number of ears in a listener's head. We can shed light on this by describing the *environmental role* of a loudspeaker as an interpolation within a structural triangle with the following corners:

- The audio channel projected from a single loudspeaker is a stand-in for an *environmental stream*.
- The audio channel is directly connected to one of the ears of the participant as a *sensory channel*, e.g. by headphone.
- The channel is part of a multi-channel array to be projected from loudspeakers that are each heard by both ears. Spatial impressions are encoded in inter-channel signals.

We find the first case realized for example by the projection of film dialog from the center channel in order to constrain the localization of the actor's voices to the center of the screen. The second case conforms to the binaural application of sound to the listener's ears via headphones, and the third case is found in all loudspeaker arrays that surpass the two stereo channels in number, such as the cinema and home-theater audio formats promoted by the movie industry (5.1, 7.1, 9.1, et cetera). and finds its most extreme realization in wavefield synthesis systems in which a single loudspeaker is never heard as a discrete *sound-source* on its own and instead always appears as a contributing element in the synthetic creation of an environmental sound field. A more detailed investigation into the relationship environmental streams, audio channels and the sensory array of the participant needs to be topic of a future

publication.

1.2. The Auditory Scene: Stream formation and selection or *perception-as-action*?

The process by which acoustic energy that arrives at the ear is transformed into auditory experience is the concern of psycho-acoustics research. The description of principles and processes involved in the formation of objects and streams in the perception of time-based content can be approached from a variety of perspectives. A very influential school of thought in the area of perceptual object formation are the *Gestalt Principles of Perception*, a set of rules and tendencies that seem to underlie our structural interpretation of the environment—the emergence of forms, boundaries, shapes, foregrounds and backgrounds and so forth [3]. While Gestalt Psychology has its origin and focus in the analysis and description of *visual perception*, we can interpret A.Bregman's well known work on *Auditory Scene Analysis* as a correlate for auditory domain [1]. Similar to the grouping principles of gestalt psychology, Bregman sees auditory perception as a process of fusion and segregation that results from properties and features of the acoustic signal: On the one hand the fusion of perceptual elements depending on their spectro-temporal structure (harmonicity, common onset/offset, common fate in the frequency or amplitude domain, et cetera), and on the other hand the linking of distinct events into perceptual streams depending on their similarity in auditory *feature-spaces*: For example, the distinct timbre- and pitch-spaces of a flutes, violins, birds and cars cause them to segregate into distinct perceptual objects and continuous perceptual streams. Here, spatial location is one factor among others.

It has been argued that the role of the *perceptual object* is not sufficiently described as a bottom-up coagulation juxtaposed to the process of attentional selection, but that there exists an important infusion of low-level stream segregation by cognitive processing, and that the *objects of perception* can in fact simultaneously be regarded as a basic unit of both cognition and attention [4]. In the psycho-acoustic domain these relationships are being investigated in the work of B.Shinn-Cunningham [2].

Another approach to the structural interpretation of perception occurs in the wake of the theory of environmental perception established by J.J.Gibson [5]. Gibson avoids the bottom-up and top-down structures of gestalt theory and instead sees perception as a *direct* process that dispenses with the differentiation between the stimulus, the environment and its perception. Alva Noë in turn interprets this direct perception *as action*—the involvement of the participant's body in a direct performance of perceptual enactment [6].

From these diverse backgrounds, we can consider the segregation of perceptual objects, streams and behaviors that are available to selection by focus and attention not only as the outcome of a feature-based coagulation, but also as inference of patterns and expectations by the observer and finally, following Noë, the activation and involvement of specific *perceptual strategies*: In the context of this presentation, we would like to address this conceptual fusion between the formation of *perceptual streams and objects* and the involved strategies of it active perception as the an outward perceptual activity of *modal streaming* that is performed by participants. Perceptual involvement with media displays can be regarded as an application of modal strategies by which participants discover, approach and become involved with the environment. Modal streams are distinct from *sensory streams* as they can alternatively span multiple sensory modalities or become segregated within a single sensory stream—but also in distinction from *perceptual streams* that emerge from a bottom-up fusion of sensory stimuli. What we mean by *modal streams* is the performance of a perceptual strategy by the perceiving participant in a continuous process of active perception in the senses of Noë —a perceptual involvement the participant might be unaware of [6]: Both the conscious effort of looking up a youtube video and involuntary eye movements in the observation of an image can be regarded as aspects of a *modal strategy of active perception*.

1.3. The simultaneous manifold

In audio-visual media, perceptual objects and streams can span multiple sensory modalities: A car driving by, people talking in the background, a record player playing diegetic (in-scene) music, et cetera. We experience independent simultaneous multi-modal objects that form relationships and groupings, a whole that consists of simultaneous parts: Our experience of a time-based media artifact could be described as a *differentiable simultaneous manifold*.

As we attend the multiple seemingly independent entities that occur in juxtaposition, superposition and sequence within the mediated content, we tend to become oblivious to the technological transmission channels or the way the media system addresses our sensory channels we have described in 1.1. And instead become immersed in a mobile panorama of perceptual objects and streams that is at the same time *coherent* and *navigable*.

While the strict definition of attention allows the perceptual selection of only a single object or stream [2], the perceptual simultaneity of distinct but coherent perceptual streams we encounter in auditory media suggest that the *shape of what we can attend to simultaneously* is wider than a single *perceptual object* or *auditory stream* in the definitions of Bregman and Koehler.

Evidently, our potential for simultaneous perception is characterized by limitations. Barbara Shinn-Cunningham describes the middle-ground between perceptual object formation and attentional selection as a *biased competition* that is decided either by the volition and attentional direction of the perceiver or the salience of the perceptual object. Following the idea of perception as combination of simultaneously activated *modal strategies*, we may describe these potentials for simultaneous perception as a repository of perceptual resources that is available to the observer.

2. PERCEPTUAL STREAMS AS PERSISTENT PERCEPTUAL INTERFACE

Auditory streams in the sense of Bregman are characterized by a dichotomy of *mobility* and *persistence*: On the one hand, the stream itself persists over time and is attributed to or accountable for the emergence of persistent objects within our environment. On the other hand, its appearance can change and modulate, and its variability has the potential to encode information within itself: A speaking voice, figuring prominently in the famous auditory scene example of the *cocktail-party* [7], is characterized by a persistence that allows the party guest to navigate the auditory scene with their attentional focus. But the interior, the *content* of the stream is characterized by variability: What is being talked about, how it is being said, the specific sounds of vowels, consonants, phonemes, how the physiological performance of the speaker contextualize the individual voice, et cetera: The modal stream can be

interpreted as an *interface* that allows the discovery of previously unknown aspects and properties of the environment. Upon closer inspection, streams can in turn disintegrate into a manifold of independently observable features: Streams within streams, accessible within one another through progressive attentional disclosure as it was described for example in Merleau-Ponty's phenomenological analysis of perception [8].

As a *perceptual interface* toward our environment, modal streams provide us with an access of relative persistence through which we provide attention to environmental processes. In this way, we can see them as a bidirectional relationship: On the one hand, they form a channel through which environmental information reaches us, on the other hand, a pre-set strategy to interpret the environment is already implied in the establishment of the stream itself.

3. APPROACH FROM INSIDE: PERCEPTUAL RESOURCES

Multiple streams can be present in our environment simultaneously, but often we can not attend all of them at the same time: We see ourselves surrounded by opportunities to involve our perception and action, but we can only realize a very limited subset of them at any given time. In cognitive science, we find this formalized as a juxtaposition between an array of disclosed perceptual objects and streams on the one hand and the process of our shifting attentional selection on the other hand [9, 2].

However, we need to acknowledge that in the pre-attentional formation of perceptual objects the a *type* of object is already defined, and moreover, these different *phenomenological types* of streams are characterized by a different potential to be attended simultaneously. More than a general *sensitivity for sound waves*, hearing involves an a priori *listening-for*, a perceptual top-down pre-organization, and it appears to be that different types of listening engagement are characterized by a varying potential for simultaneity, to be occurring in parallel or at the same time as other engagements.

For example, it seems evident that we only have the potential to fully engage and understand a single stream of type *speech*. Multiple simultaneous language streams will lead to a discrimination of the streams into *attended* and *peripherally attended* speech—or, if that is not possible, confusion and unintelligibility are the consequence. We find an even more extreme case in music, in which the addition of a second music stream into the environment leads to an effective destruction of the music with only very limited potential to selectively attend one of the coinciding streams. Then again, we seem to be able to let multiple different non-speech environmental sounds occur simultaneously without a similar destructive effect. In a structural analysis of these relationships, we can distinguish the following cases:

3.1. The navigable multiple and polyphony

3.1.1. Navigable multiple

As we can see in the example of the cocktail party, perceptual streams can form a *navigable multiple*: While not all streams can be attended simultaneously, the streams are still accessible to participant's select and engagement. We can only attend to one conversation at a time, but which one is up to our attentional navigation of the auditory scene.

3.1.2. Parallel simultaneity and polyphony

In certain cases, modal streams can become accessible in parallel simultaneity: We can experience a collection of streams in simultaneous superposition while they still retain their own identity and potential for an increase of depth of attention. We can see an example in the potential of speech and music to be present simultaneously—as opposed to the superposition of two *musics* or the presence of two speaking voices simultaneously which is immediately characterized by conflict. We can compare this to *musical polyphony* which represents another example: In a 4-part fugue, the voices retain independence to an approach of analytic listening, but cohere to form an aggregate: Attentive selection may shift between focusing on a single stream or the global perception of the harmonic relationships resulting from their combination. The layers of a movie soundtrack can be seen as another example: Each of the layers of the soundtrack—dialog, music and the various sound effects—is characterized independence that allows them to be created by different production teams, can reside in a different phenomenological area as Michel Chion describes in his book *Audio-Vision*[10]. Nevertheless, a coherent experience is created that has the potential to subsume the individual constituents within it. In contrast to the *navigable multiple* from which the participant can freely pick streams to attend, we can call this case in which distinct streams form a new coherent whole the *polyphonic multiple*.

But next to the formation of navigable and polyphonic manifolds, perceptual objects and streams can also merge or obstruct each other.

3.2. Correlative merge

If modal streams contain correlated behaviors this may result in their perceptual fusion into a single more complex stream or group of connected developments. This is the case for example for complex sound objects or audio-visual coherence in the context of cinema sound (for example, a car drive-by).

It is important to note that while correlative effects occur within our perceptual environment, for example the micro-correlation between a sound source and its reflection that leads to the encapsulation of the reflection into the *spatial timbre* of the sound source, correlation can also be discovered as an effect of self-motion: We may hypothesize that the impression of spatial persistence, for example of architecture, could be interpreted as an effect of correlation between the self-motion of a participant and the perceptual change in the appearance of the architectural environment. The merging of perceptual elements that show correlated behavior is in accordance with the rules governing the perceptual fusion and segregation of streams [1].

3.3. Destructive merge

The destructive merge is an everyday experience: Streams mingle together and overlap making each other mutually indistinguishable, comparable to two layers of handwriting written in top of each other. For example the projection of two speaking voices from the same loudspeaker, or the simultaneous presence of two violin sonatas usually lead to a destructive merge of the simultaneous streams.

In the hierarchical perspective of bottom-up and top-down formation of perceptual objects, the mutual obstruction of perceptual

objects and streams can occur on any level of formation or attentional selection—from *energetic masking* in the sensory channel to various effects of *informational masking* or failure in attentional selection [2]. Coming from the perspective of direct perception, we can describe the mutual merging and masking of modal streams as *perceptual resource conflict*. Like the navigable and polyphonic manifold, we can interpret merging and masking as a structural dependence and relationship between the perceptual resources that we apply to different aspects of the environment over time.

4. INTERLUDE1: PITCH, SPECTRAL MORPHOLOGY AND THE MODAL STRATEGY OF MELODIC LISTENING

A popular example of perceptual fusion is the phenomenon of instrumental timbre. As we know, the perception of timbre is related to the amplitude and phase relationships of partial frequencies that are connected by a *common fate* in frequency and amplitude. Preferably, the partial frequencies have *harmonic* ratios [11].

But beyond the emergence of *pitch* and *timbre* as independent categories, we might say that to hear a sound as a musical note, as an element within the context of a melody, is more than just an effect that emerges from a partial relationship within the signal itself. Music implies a self-application of the participant to the melody through a strategy of *melodic listening*. What we mean by that is exemplified in the *speech-to-song* illusion described by Diana Deutsch[12]: A repeated fragment of spoken word is initially approached with a strategy of *speech listening*. Upon multiple repetition, the strategy shifts, and what is heard becomes more and more a sung melody. The signal has stayed the same, what has moved is the listener. We can say that the strategy of *melodic listening* we apply to music in fact determines our attitude and thereby our interpretation of the music.

In the opposite direction, we can also find musical examples in which our—intuitive or trained—strategies of *melodic listening* have been intentionally subverted: If the harmonicity of the spectrum or the common fate of the partials is disturbed, the fusion into a sound characterized by a single pitch and timbre can break up and begin to sound bell-like: We may hear *multiple simultaneous pitches* within a single sound, especially if we have trained ourselves to navigate such frequency mixtures. If furthermore the common fate of the partials is disturbed, the experience of the sound can split up into even more independent entities all together.

A music piece in which these effects can be experienced in an exemplary way is Karlheinz Stockhausen's piece *Cosmic Pulses*in which sound layers, clearly delineated by a common fate in the area of frequency, amplitude, spatialization, develop interior worlds due to the inharmonic *split* spectra and the micromodulations within the spectral composition of each layer: An unsettling experience as we find our modal approach to the hearing of sound constantly challenged and on the edge of disintegration, all the while new layers are piled atop one another [13]. In his own words, Stockhausen admits that one might not be able to attend all contained streams during one individual listening run:

> If it is possible to hear everything, I do not yet know– it depends on how often one can experience an 8-channel performance. In any case, the experiment is extremely fascinating! [14]

5. PERCEPTUAL RESOURCES: LISTENING AS SELF-APPLICATION

We often find music tracks organized into a *playlist*, the reason being that we are generally unable to appreciate two musics playing simultaneously—we prefer to attend them in sequence. When we superimpose two *musics*, they usually do not combine *navigable multiple*. While details of each music track remain accessible to attentive selection, others merge into a combined perception that appears not so much a summation of its parts but a different experience in itself. We may pick up on familiar instrumental timbres, vocalists, melodic fragments and recognizable moments of each music even when it is superimposed with another music, but certain aspects become very hard or even impossible to perceive when presented in temporal coincidence. To pull it down to a common sense statement: Music is a time-based art and lives from the fact that elements are presented in succession, with specific duration, intensity—and the attentive presence of the listener.

While simultaneous melodic lines for example can add up to a navigable polyphony—whether this occurs in the confines of musical meter and harmonic counterpoint as in Bach's music or as a stochastic and *chaotic* process one such as in Xenakis or Ligeti shall be another question—but it appears that only one sense of *harmony* or *tonality* seems to be possible at any moment: If multiple harmonies coincide, we do not hear both at the same time. In the case of harmony, we also have difficulties to listen to them as navigable parallel presence in the same way that we might attend to two talkers at a cocktail party. What emerges is a new *bi-tonal* harmony—a new tonality in itself.

We can find a similar behavior in the perception of rhythm. If two different repetitive rhythmic structures coincide, we seem to be unable to hear them as two separate rhythms at the same time. In some cases they might form a navigable multiple if they can be attributed to different modal streams, but more often they will combine into a new rhythmic structure. Even while we might be able to discern what meter each music piece is by selectively attending to individual instrument timbres if one of the coinciding *musics* is characterized by repetitive patterns, the overall impression of the rhythm will be lost.

The phenomena of *harmony* and *rhythm* contain phenomenological aspects that resist the formation of a *navigable multiple* or even a *polyphonic multiple*. We can describe them as *perceptual resources*: A limited potential to simultaneously attend to environmental phenomena. The musical features of *harmony* and *rhythm* are akin to our ability to only attend to one language stream at any given time, albeit with different structural demands on simultaneity and another navigation strategy for the participant. While cocktail parties encourage a manifold of simultaneous conversations, there usually is only a single music track playing in the room. Our listening can handle a coincidence of rhythm, harmony and an environment of navigable conversations, but not two incoherent harmonies and rhythms. [1]

[1]The first modern composer to exploit the collision of different harmonies and rhythm was arguable Charles Ives who is known for experimenting with marching bands performing pieces of different harmony and rhythm while marching through his home town—an experience he would later emulate in the polymetric sections of his symphonies.

Figure 1: You can shift between seeing an old or young woman in this famous image [15]. However, it appears problematic to see both at the same time.

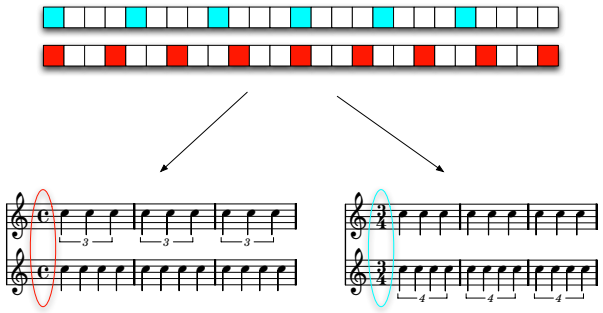


Figure 2: You can shift between hearing this time sequence as 4/3 or 3/4 polyrhythm—as two distinct rhythms occurring in 4/4 or 3/4 time respectively.

6. POTENTIAL ORIGINS OF MODAL STRATEGIES

The different morphology of the modal strategies involved in the perception of speech and music begs the question what origin they can be traced back to.

Of course we have to assume that the establishment of *see-as* and *listen-for* patterns that underlie these phenomena is subject to continuous improvisational adaptation, optimization and intuitive experimentation. Our taste in music changes, as does our perspective on all other aspects of life. One way to describe this open epistemological field is the area of the *cognitive body* I have described in [16]. However, we could for example list three potential channels through which modal strategies could emerge: *Learning and experience, evolutionary development and emergence*.

Evolutionary we can assume that basic modal strategies are made available to us through an expression of our genes. For example, our basic sense of hearing—the potential to perceive sound in general can be attributed to the fact that we have ears which evidently evolved through natural selection. In this area there are also the physiological and neuro-physiological properties of our body that can become an active element in the task of perceiving sound—for example the experience of groove. In his book *Sweet Anticipation*, David Huron traces musical experience back to the evolutionary history of auditory processing the central nervous system [17].

Emergent modalities address us from a stream of perceptual events that enters our perception from our environment: Something catches our attention without a clear pre-formed interpretation or expectation: There is an a-priori sense and experience of *potential meaning* in the experience of the signal, motivating a process of attentional observation which leads to the accumulation of hypotheses, inferred persistencies like patterns, objects and agencies: The self-organizing emergent collection of assumed and expected underlying behaviors. This can immediately be observed in the process of *listening to music*.

A *learned* modality can be seen in the ability to attend speech: While we might be endowed with an innate, potentially *physiologically pre-disposed* [18] tendency to attribute meaning to reoccurring sound patterns, the specific language we speak comes toward us from the environment we grow up in—the interactions we have as children with our environment. We might say the speech channel emerges in a self-driving process of improvisatory rehearsal by a continued contribution of trial, error, conscious effort in production and attention.

7. INTERLUDE2: POLYRHYTHMS AND THE SHIFT OF PERSPECTIVE AS PERCEPTUAL SELF-APPLICATION

The strategies by which we listen to our environment are characterized by a degree of conscious control. We can see this in the case of polyrhythm perception. The perception of polyrhythms is split into the perception of a *primary beat* that conforms to the perceived *meter* of the rhythmic structure, and a *secondary beat* which is heard as being offset or as "standing against" the primary beat. While the temporal structure of the events themselves stay identical, listeners have the potential to consciously navigate between different listening perspectives on the polyrhythm by applying the modal strategy of the *meter* to each of the two layers, shifting the way the polyrhythmic stream of beats. We can compare this process to way ambiguous images appear, for example the famous picture that can be seen as an old or a young woman, depending on the way we apply our strategy of *seeing a face*. In both cases, we can not take both perspectives at the same time.

8. MUSIC, SPEECH, THE *NATURAL ENVIRONMENT* AND SONIFICATION: DISTINCT MODAL STRATEGIES

Taking a closer look at the activity of listening to music, speech and sounds from the natural environment, we can distinguish different relationship of the activity and the participant: We find *modal strategies* in the interpretation, approach, following and tracking of the sound and what is encoded within it that imply a different kind of involvement.

8.1. Environmental sound

When we are immersed in natural sound scenes, we are experiencing sounds in their natural state, as an *identity of the sound with its source*. Unlike speech and music, which are strategies used by human beings to target the perception of other human beings in order to achieve a specific effect, the sound caused by the wind in our ears is a property of the air and the wind. Animal sounds are an aspect of the animal. The presence of water is announced by its spe-

cific look as well as its characteristic sounds, et cetera. Of course it has been argued that the perceptual approach toward our *natural* environment has been developed and optimized in the process of evolution, and a perceptual theory that underlines this identity of perception and the environment can be found in J.J.Gibson's work on environmental perception [5]. From this perspective, musical listening tends to appear as a secondary category—a *cheesecake of the mind*[19], and speech listening becomes yet another even more extraordinary involvement.

8.2. Music

Music is generally expected *to produce a desired effect by itself*, without any analytical effort of the participant. What we hear is not experienced as property of the external environment, but an emotion, meter, rhythm, melody, et cetera, that emerges within an inherently *human way of listening*. Arguably, listening to music is not an involvement with the outside world but in fact with our own potentials of having an aesthetic experience. In order for music to appear, the participant has to provide specific perceptual resources—for example what we have previously circumscribed as the potentials for *harmonic* and *rhythmic listening* or the potential to experience sublime emotions as laid out by David Huron [17]. We could describe the musical experience as a *massage* of these resources, and the participant has little more to contribute than to remove potential distractions from the environment to make sure nothing else will occupy the required perceptual potentials and thereby *mask* and *occlude* the musical experience. As we accumulate experience throughout our lives, new perceptual resources form, and our taste of music changes: We can continuously discover new and interesting aspects in music, however, when the music *doesn't work*, when it causes dissatisfaction or confusion, we usually do not blame ourselves: The composer, the performer, the sound engineer or the home stereo is at fault, while our ability to listen to and enjoy music is often considered an innate aspect of our humanity.

8.3. Speech

Speech on the other hand is very obviously an *acquired* perceptual strategy. We are not born with the language that our parents speak, and we have to learn both the production of speech as well as its understanding: Native language is acquired through attention, rehearsal, repetition, optimization, reflection, trial-and-error, adaptation, et cetera. Listening to language is evidently the involvement of a specific learned resource of the participant: We can only do it for one speaker at a time. In speech, the difference between the transmission channel and its content becomes evident: The fact that a person is talking is to a large degree independent of what they are going to say. The involvement of decoding language has a degree of independence from the circumstances the language is heard in—even though we take the situation of what is being said into account.

8.4. Sonification

When we interpret sonification not only as a strategy to organize, create and render sound, but inversely as a *modal listening strategy* or, to put it simpler, a *way of listening*, we can see how it is different from environmental sounds, speech and also music:

In comparison to *natural environmental listening*, sonification necessarily has to communicate its data by using properties of sounds that are *inherently detached from their source*. As such sonification is comparable to a learned listening strategy like language. It is designed to target our perceptual potentials in a specific way, but in order to *encode something other than itself* in a similar way speech or a technological media channel would.

This involvement of the listener *to see something in the sound which is not itself* is also a difference between sonification and music. To Paul Vicker's dichotomy of *sonification concrete* or *sonification abstraite*[20] I would like to add that it is not sufficient to place the accountability for the appearance of sound into the human strategy for sound/music-generation alone. This would be comparable to placing the accountability for the meaning of speech only into the act of speaking while disregarding the involvement of *understanding*.

When we listen to Xenakis, John Cage and Alvin Lucier, we may indeed hear something that is comparable to *sonification heard as music*. The use of data appears as an element subverting the continuum of intentionality that is seen to reach from the composer to the experience of the music listener in order to evoke *open potential* in the participating listeners can be seen in the context of a larger cultural context of this era, as outlined by Umberto Eco's idea of the *Open Work* [21]. A further superficial kinship is generated in the sense of *unfamiliarity* and potentially *initial discomfort* that results from the fact that this strategy of *New Music* and sonification require ways of listening that are unfamiliar to the listeners of speech, natural environments and pre-20ieth century music.

But it is evident that the relationship between the sound and the listener as well as within the listener's involvement is very distinct: In the first case, a composer is exploring a strategy of generating an *aesthetic experience within the sound and its performance itself* that appears as new and unfamiliar to the listener. The plan is to invoke the curiosity of the listener and tap into our innate tendency to react to new experiences in our environment with the development of a complementary listening strategy: We always want to make sense of the world of course, we want to know what's going on, so we reach out and gather around what we do not understand.

The end state of successful sonification however is that the sound, or any aspects of a musical experience in fact *vanish from the listeners perception*, and what shines up behind the auditory transmission of information are the data that underlie the sonification: The listener is not consciously involved in listening to sound, but becomes connected to the data and relationships that are encoded within it, in a similar way that the listener of speech become oblivious to the sound of phonemes, and the pitch of the voice, and instead focuses on *what is being said*—a process we saw reversed in Deutsch's Speech-To-Song Illusion [12].

The sound features become an intermediate encoding step in the communication of data, and the experience is mediated by music, but in the end primarily non-musical: The difference between *message* and *massage* in the sense of McLuhan [22]. Whether the sounds embodied in this process are derived from sound-making properties of our natural environment or electro-acoustic *acusmatic* sound that has no other source than a loudspeaker [10], or whether the sound properties share a kinship to *musique concrete* or tonal music—even whether the sound is comfortable, aesthetically pleasing, beautiful et cetera—become secondary criteria similar to whether the sound of the announcer's voice on the train platform is pleasant to listen to.

That being said, evidently *New Music* has shown is the way of opening up musical accountability to *non-intentional* elements such as data values and thereby created a bridge for listeners to

open their ear to the qualities of *sounds detached from their cause*, and this achievement is of course interesting to acknowledge from the perspective of sonification. In a previous publication we have argued that referential sound, for example the famous use of *piano samples* as carriers of pitch information, can lead to a loss in perceptual detail—the technological transformations that lead to the formulation of *musique concrete* have shown us the way how to *listen to spectral qualities* of sound and thereby made a new perceptual approach possible. In this sense, we might indeed be able to *let music shows us the way*, but the focus has to be the activity of the listener and participant.

What makes the world behind the sound appear is the listening strategy of the participant, the artist and composer ideally becomes as invisible as the *designer of a language*.

9. SUMMARY: DISCOVERING THE MODAL STRATEGIES OF SONIFICATION THROUGH THEIR *POTENTIALS FOR SIMULTANEITY*

I derived the concept of *modal strategy* from a structural description of our potential to appreciate simultaneous multitudes of specific kinds of processes in our environment—speech, harmony, rhythm are three examples. From this position I argued that listening is characterized by specific potentials for simultaneity that are inherent in the perceptual approach toward our surroundings, for example the ones listed in 3.1.

From here we may ask: What needs to be *moved out of the way* if a sonification strategy should be perceived successfully? Do sonification strategies allow to be perceived simultaneously (like music and sound effects), or do they mask each other? What is the specific domain the competition, collision or masking occurs in—is the masking *energetic*, *informational*, or inherent in the the activity of *participation*, such as attentional selection, focus, following and other aspects of *perception-as-action*? Under what circumstances can a sonification strategy generate a *navigable multiple* or *polyphony*?

I expect that an inquiry from this participant-centric perspective will in fact lead to more successful sonification designs that, insted of placing the accountability into the mappings and modals of data are motivated by a participant-oriented interest in *auditory scene synthesis*—a line of work that is already in process in the developments of stream-based sonification [23].

Through the development implementation and application of new modal listening strategies sonification can become an auditory interface that allow the active involvement of the participant, enabling them to experience accountable structures and perceptual properties far beyond an experience of *sound modulated by data*.

10. REFERENCES

[1] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*. The MIT Press, Sept. 1994.

[2] B. G. Shinn-Cunningham, "Object-based auditory and visual attention," *Trends in Cognitive Sciences*, vol. 12, no. 5, pp. 182 – 186, 2008. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1364661308000600

[3] W. Köhler, *Gestalt Psychology*. Liveright, New York, 1947.

[4] J. Feldman, "What is a visual object?" *Trends in Cognitive Sciences*, vol. 7, no. 6, pp. 252 – 256, 2003. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1364661303001116

[5] J. J. Gibson, *The Ecological Approach To Visual Perception*, new edition ed. Psychology Press, Sept. 1986.

[6] A. Noe, *Action in Perception*. The MIT Press, Mar. 2006.

[7] E. C. Cherry, "Some experiments on the recognition of speech, with one and with two ears," *The Journal of the Acoustical Society of America*, vol. 25, no. 5, pp. 975–979, 1953. [Online]. Available: http://link.aip.org/link/?JAS/25/975/1

[8] M. Merleau-Ponty, *Phenomenology of Perception*, 2nd ed. Routledge, May 2002.

[9] H. E. Pashler, *The Psychology of Attention*. Cambridge, Mass: MIT Press, 1998.

[10] M. Chion, C. Gorbman, and W. Murch, *Audio-Vision*. Columbia University Press, Apr. 1994.

[11] C. Plack, A. Oxenham, and R. Fay, *Pitch: neural coding and perception*, ser. Springer handbook of auditory research. Springer, 2005.

[12] D. Deutsch, T. Henthorn, and R. Lapidis, "Illusory transformation from speech to song," *The Journal of the Acoustical Society of America*, vol. 129, no. 4, p. 2245, 2011. [Online]. Available: http://asadl.org/jasa/resource/1/jasman/v129/i4/p2245_s1

[13] N. Collins, "Karlheinz stockhausen: Cosmic pulses," *Computer Music Journal*, vol. 32, no. 1, pp. 88–91, 2008. [Online]. Available: http://dx.doi.org/10.1162/comj.2008.32.1.88

[14] K. Stockhausen, "Cosmic pulses," CD Liner Notes, Kürten: Stockhausen-Verlag, 2007.

[15] W. E. Hill, "My wife and my mother-in-law. they are both in this picture - find them," in *Puck*. Washington, D.C. 20540: Library of Congress Prints and Photographs Division, 1915, vol. 78, no. 2018, p. 11. [Online]. Available: http://www.loc.gov/pictures/resource/ds.00175/

[16] J. Gossmann, "From metaphor to medium: Sonification as extension of our body," E. Brazil, Ed., International Community for Auditory Display. Washington, D.C., USA: International Community for Auditory Display, June 9-15 2010. [Online]. Available: http://icad.org/Proceedings/2010/Gossmann2010.pdf

[17] D. Huron, *Sweet Anticipation: Music and the Psychology of Expectation*. The MIT Press, Mar. 2008.

[18] C. P., "Temporal codes, timing nets, and music perception," *Journal of New Music Research*, vol. 30, pp. 107–135, June 2001. [Online]. Available: http://www.ingentaconnect.com/content/routledg/jnmr/2001/00000030/00000002/art00002

[19] S. Pinker, *How the Mind Works*. W. W. Norton & Company, Jan. 1999.

[20] P. Vickers and B. Hogg, "Sonification abstraite/sonification concrete: An 'aesthetic persepctive space' for classifying auditory displays in the ars musica domain," C. F. A. D. N. E. Tony Stockman, Louise Valgerur Nickerson and D. Brock, Eds., Department of Computer Science, Queen Mary, University of London, UK. London, UK:

Department of Computer Science, Queen Mary, University of London, UK, 2006, pp. 210–216. [Online]. Available: Proceedings/2006/VickersHogg2006.pdf

[21] U. Eco, *The Open Work*, 2nd ed. Harvard University Press, Apr. 1989.

[22] M. McLuhan and Q. Fiore, *The Medium is the Massage*. Gingko Press, Oct. 2005.

[23] S. Barrass and V. Best, "Stream-based sonification diagrams," Paris, France, 2008, inproceedings. [Online]. Available: Proceedings/2008/BarrassBest2008.pdf

DEMONSTRATION OF AN OUTDOOR AUDIO SHOOTING GALLERY

Mark A. Ericson

Army Research Laboratory,
520 Mulberry Point Road,
Aberdeen Proving Ground, MD 21005
mark.a.ericson.civ@mail.mil

Matthew N. Vella

Army Research Laboratory,
520 Mulberry Point Road,
Aberdeen Proving Ground, MD 21005
matthew.n.vella.civ@mail.mil

ABSTRACT

An audio shooting gallery was created to demonstrate the immersive and interactive audio capabilities of the Army Research Laboratory's Environment for Auditory Research. The demonstration participants come from a wide variety of backgrounds including students, Soldiers, scientists, and technical managers. Targets, selected by the shooter, include various wild animal vocalizations, bird calls and animal rustling sounds. The shooter also selects a weapon which is appropriate for that particular target. Audio targets simulating the wild animal being hunted are played over random loudspeakers in the 110 meters by 25 meter outdoor range. A nocturnal hunting simulation can also be created indoors in the Distance Hall by controlling the ambient light level. As the shooter engages the targets, weapon-specific sounds are played through loudspeakers surrounding the listener to enhance the immersive experience. Environmental sounds, appropriate for the selected wild animal target, are played over the entire outdoor loudspeaker array. Distracting, non-target sounds are played from other loudspeaker locations to force the listener to identify and discriminate the correct audio target among distracting noises. Hits, misses, and false alarms are scored and displayed to the shooter to provide accuracy and performance feedback.

1. INTRODUCTION

A demonstration of the audio capabilities of the Environment for Auditory Research (EAR) [1] was created, based roughly on a shooting gallery amusement park game [2]. In the audio only gaming demonstration, the shooter selects a wild animal sound for the target and a weapon sound of choice for hunting that animal. Based on the shooter's choices, the computer program automatically selects appropriate ambient environmental sounds and other non-target animal sounds as distracters. Up to eleven distracting sounds and one target sound can be played simultaneously over loudspeakers in the outdoor space, and up to sixteen total sounds in the indoor space. When the shooter has identified the location of the target animal sound, he or she aims a pistol shaped pointer at that location and presses the trigger button. Appropriate near-field weapon sounds are played from loudspeakers surrounding the shooter. Also, hit or miss type sounds are played from downrange loudspeakers immediately after the weapon firing sound. Bullet impact and ricochet sounds are played based on the direction the gun was aimed when the target was missed. Target impact sounds are played when the weapon was aimed within an appropriate radius for a particular target. This interactive demonstration provides an amusing way to show the spatial sound generating capabilities of the EAR. The outdoor environment helps to immerse the participant in a large outdoor environment in which realistic animal and habitat sounds can be presented, creating a plausible space in which daytime hunting could occur. The indoor space of the Distance Hall creates a believable night-time shooting experience.

2. FACILITY DESCRIPTION

The Open EAR is an outdoor research space for playing sounds in natural auditory environments. A picture of the Open EAR from inside the doorway of the EAR's Distance Hall is shown in Figure 1. Audible environmental sounds include wind, bird calls, distant aircraft and ground vehicles, and occasional real explosions from Aberdeen Proving Ground's test range. Nominal background noise levels range from 10 to 50 dB SPL.

Figure 1: The Open EAR instrumented with eight loudspeakers as seen through the middle doorway of the Distance Hall.

The Open EAR is instrumented with eight audio interface units, which include loudspeaker, microphone, and 120V AC power outlets. Up to twelve loudspeakers and twelve microphones can be placed at any location spanning the outdoor space measuring 110 meters long by 25 meters wide. All of the presented audio signals are generated by a single computer inside the facility's main control room.

3. DEMONSTRATION

The task involves the correct identification, accurate aiming, and simulated shooting of various target sounds somewhere in the outdoor space of the Open EAR or the indoor space of the Distance Hall. Simulated target sounds "pop up" among a din of background noises, indigenous to the habitat of various wild animal targets. The immersive virtual audio environment of the EAR facility creates a truly first person gaming perspective. The specific equipment and procedures involved in the demonstration is described in the following paragraphs.

3.1. Equipment

The demonstration utilizes: 1) the Open EAR, 2) loudspeakers, 3) an Intersense IC2+ InertiaCube sensor [3] inside a gun-shaped pointer, and 4) recorded sounds of animals, environments, and weapons fire [4][5][6]. A shooter using the gun-shaped pointer aimed at an outdoor target is shown in Figure 2.

Figure 2: A gaming shooter aiming at an acoustic target in the Open EAR.

Three sets of powered loudspeakers are available for use in the Open EAR. These include Genelec 8030A studio monitors and 7060B active subwoofers, JBL PRX512M, and Meyer Sound MM-4XP loudspeakers. Loudspeaker selection is based on the lowest frequency content of the target sound. Sound source locations are sometimes at the physical loudspeaker location and sometimes between the physical loudspeaker locations. Phantom audio sources between loudspeakers are created by using a vector-based amplitude panning (VBAP) algorithm [7]. Stationary and simulated moving sounds can be created using this algorithm.

3.2. Procedures

The subject chooses options from the computer's graphical user interface. Choices include type of wild game to be hunted, audio environments of an animal's habitat, hunting weapons, and skill level of the wild game hunter.

Wild game options include several types of typical wild animals. The bird option includes pheasant, wild turkey, and quail. The big game option includes elk, wolves, and deer. The big cat option includes lions, mountain lions, and tigers. The reptile option includes alligators and rattlesnakes. These last two options, big cat and reptile, include animals that can strike back at the gamer. After a miss, the attacking predator sound appears closer to the shooter. After two misses of an attack animal's target sound presentation, a ferocious growl is played over the Distance Hall loudspeakers surrounding the gamer, followed by a succulent chomping sound. The demonstration is abruptly ended and a "game over" message is displayed on the computer screen.

Audio environments include several typical soundscapes in which a chosen animal is typically found. Examples include a mountainous terrain, nocturnal woodlands, an open meadow, and an African jungle. The mountainous terrain and open meadow soundscapes were made from recorded soundtracks of real environments. The nocturnal woodlands soundscape was created by mixing selected distracting animal sounds with a real recording of a night-time woodland. Likewise, the safari soundscape was created by mixing distracting, non-target African jungle animal sounds with a background jungle waterfall recording.

Hunting weapons include archery, small arms, semi-automatic long barreled guns, fully automatic firearms, and wide area explosive devices. The archery weapon simulates the launching sound of either a bow and arrow or a crossbow. Small arms weapons include pistols and side-arms. Long barreled guns include a standard .22 rifle, a military style M-1 rifle, and a double barrel shotgun. Multiple round weapons include M-16, AK-47, and .50 caliber automatic weapons. Wide area explosives include grenades and small diameter bombs. Lastly, a thermonuclear device can be chosen as a weapon of last resort by pressing the top button on the gun-shaped pointer. However, detonating such a weapon usually destroys the wild animal, the gamer, observers, the surrounding environment in a 20 mile radius, and quickly ends the demonstration.

The shooter can select one of three skill levels from the GUI interface. The levels include a novice hunter exemplified by Elmer Fudd, a weekend sportsman, and a professional marksman characterized by Clint Eastwood. A novice scores a hit when the weapon is aimed within 3 degrees of the target. A sportsman scores a hit for better than 2 degrees of accuracy. A marksman requires 1 degree of accuracy to get credit for a

direct hit. When a target is hit, an impact sound is played over the target loudspeaker for that particular projectile hitting an object. When the shooter misses the target, a ricochet sound is played from the loudspeaker closest to the direction the shooter was aiming his weapon.

3.2.1 Acoustic Stimuli

Stimuli were chosen from the SFX source database, the Macauly Library, the Networks Sound Effects Catalogue, and custom recorded sounds. Some of the small arms and automatic weapons fire were recorded at the known-distance range on Aberdeen Proving Ground. Microphones were placed on a 16 meter arc around the shooter from 0 to 120 degrees off the target line.

Monaural soundscapes were recorded using a B&K 4165 microphone and a B&K 2804 preamplifier onto a Sony digital audio tape recorder. Spatial soundscapes were recorded using an eight-channel Holophone® H2 Pro head recorded through a PreSonus FireStudio FireWire recording system onto a laptop PC. Locations included rural parklands, windy forests, and Spesutie Island at Aberdeen Proving Ground.

3.2.2 Performance Scoring

Shooter performance was recorded in a signal detection framework. Hits, misses, and false alarms were calculated and displayed to the gamer after each stimulus was presented or after each shot was made. Correct hits were scored when the participant aimed the gun within three seconds from the moment the animal sound was played and within three, two, or one degree(s) of the target sound's location, depending on the expertise level of the shooter. For example, if the gamer correctly shoots a legitimate target then a hit is tabulated. A miss occurs when either the gamer does not shoot at the presented target or when the weapon is aimed and fired outside of their acceptable accuracy to the target's location. Upon completion of the shooting session, percent hit and percent miss scores are presented to the gamer. Values of d′ (sensitivity) and β (bias) score are computed, saved, and displayed on the host computer display in the control room.

4. OBSERVATIONS AND IMPRESSIONS

Feedback from participants of the audio shooting gallery is that the soundscapes appeared realistic and readily identifiable. There was general congruency between the selected animal targets and the sounds of their natural habitats. Echoes from nearby buildings added to the sense of envelopment and immersion into the soundscape simulations for loud outdoor sounds.

The ergonomics of the shooting task were acceptable. Unfortunately, the gun produced no "kick-back" or haptic force feedback to the gamer. A CO2 cartridge discharge may help to improve the feel of firing a real weapon. Delays between squeezing the trigger button and weapon sounds were very short and usually not noticeable by the gamer. The nearest target sounds were the easiest to locate and hit. Accurately shooting the far away target sounds was difficult, even at the novice skill level. This may be due to two effects. One is the relatively larger spatial image of nearby sounds. Another is that nearby targets had a high direct path to reflected energy ration, while far away targets had low direct path to reflected energy ratios. The latter condition may have caused a more diffuse sound source with a large auditory source width, which made accurate aiming to the center of the sound source difficult.

5. ACKNOWLEDGMENT

The authors thank AVI-SPL Corporation for constructing and maintaining equipment in the EAR facility. The authors thank AuSIM Corporation for the custom built gun shaped pointer with the integrated orientation sensor. The authors thank Kim Pollard for providing several hard to find animal sounds and her guidance on finding recordings of wilderness soundscapes. The authors thank Kim Fluitt for her warm weather recordings of bug-infested Spesutie Island. The authors give a special thanks to Jeremy Gaston for his close range gunshot recordings recorded around the shooters head.

6. REFERENCES

[1] Henry, P., Amrein, B., & Ericson, M., "The Environment for Auditory Research", Acoustics Today , 5(3), 2009.

[2] http://en.wikipedia.org/wiki/Shooting_Gallery

[3] http://www.intersense.com/pages/33/142/

[4] http://www.sfxsource.com/

[5] Network Sound Effects Catalogue (1992). Fourth Edition, Volumes 1-70, [CD]. Available from http://www.sound-ideas.com/network.html

[6] http://maculaylibrary.org

[7] Pulkki, V., "Virtual Sound Source Positioning Using Vector Base Amplitude Panning", J. Audio Eng. Soc. ., 45, 1997.

WERE THOSE COCONUTS OR HORSE HOOFS? VISUAL CONTEXT EFFECTS ON IDENTIFICATION AND PERCEIVED VERACITY OF EVERYDAY SOUNDS

Terri L. Bonebright

Department of Psychology
DePauw University
tbone@depauw.edu

ABSTRACT

This study examined whether visual context has an effect on the identification of everyday sounds. Scenes portraying actions that lead to everyday sounds were paired with the actual sounds, acoustically similar sounds and acoustically contrasting sounds Participants identified sounds, rated their confidence on the identifications, the veracity of the sounds and their familiarity with the sounds. Results showed that participants identified the actual and contrasting sounds correctly more often than the similar sounds, which were frequently incorrectly identified as the sound that occurred from the action in the visual scene. However, the confidence ratings for the identifications were lower for the similar sounds, and they rated them as less realistic than the actual sounds. Thus, even though similar sounds were frequently misidentified as the actual sound taking place in the scene, participants did recognize that such sounds were not quite correct for the visual action being portrayed.

1. INTRODUCTION

Watching movies is a favorite pastime for many people, most of whom readily accept the premise that the visual scene and the accompanying soundtrack, including the ambient sounds from the environment, were recorded simultaneously. In many cases, however, the visual tracks are recorded separately from the audio, and many of the sounds, especially the background noises, are recorded by producing sounds from objects other than the ones seen in the video [1]. Some of these sound effects are synthesized or sampled recordings while others are produced by Foley artists, who use a variety of different objects to produce sounds for the background sound track. The desired result is to produce a sound track that the movie viewer will perceive as realistic, regardless of what is actually used to produce a given sound. One example of a sound effect produced by Foley artists that movie watchers may be aware of is the use of halves of coconuts clapped together to create the sounds of horses galloping over the landscape. Foley artists routinely manipulate a number of objects to produce sounds for entirely different actions, such as crinkling cellophane for the sound of a fire crackling or breaking stalks of celery for the sound of bones breaking. This has led some filmmakers to argue that viewers have been conditioned by the media to expect "real" sounds that are not encountered in a natural environment [1].

Another newer application of sound effects to create a realistic experience is found in the development of virtual environments [2]. Researchers in this area have found that realistic 3-D sound environments can be produced using HRTF-constructed stimuli [3] and that synchrony between the sounds and visual stimuli is critical for realistically perceived sounds [4].

Since these examples suggest that listeners can be fooled into perceiving such sounds as realistic, it is important to determine whether people are able to correctly identify everyday sounds when they are presented without any accompanying visual stimuli. Researchers have shown that people are quite good at this in general [5, 6, 7, 8, 9], and that when they make misidentification errors, they are typically made with sounds that are acoustically similar.

Studies have also been performed to help determine if context can have an impact on everyday sound identification. Ballas and Mullins [10] and Gygi and Shafiro [11] showed that sounds embedded within a sequence help identification rates if they are semantically similar. For example, people are better at identifying the sound of a stapler if the preceding sound was a typewriter. Context has also been shown to provide enhancement for identification of visual objects within a scene [12,13,14]. However, the intermodal effects of sound and visual stimuli have not been investigated systematically in the same way. The exception to this are studies using speech that show that visual and auditory stimuli combine to produce interactive effects, such as the McGurk effect [15,16], the freezing effect [17], and the ventriloquist effect [18].

The purpose of the present study was to examine the effect of visual scenes with staged actions with objects that result in everyday sounds on the identification of those sounds. The scenes were paired with the actual sounds made by the objects, acoustically similar sounds to those made by the objects, and contrasting sounds that were acoustically dissimilar to those made by the objects. The responses collected from the participants after exposure to the sound/video combinations were identifications of the sounds, confidence ratings of those identifications, and ratings of veracity of the sounds. In addition, participants rated the familiarity of the sounds using a written list (see Table 1).

Four hypotheses were proposed based on the previously reviewed literature. First, it was expected that the visual context would affect the identifications of the sounds such that the actual and contrast sounds would be more likely to be correctly identified compared to the similar sounds. This would be the case if the acoustically similar sounds were confused with the actual sounds as suggested from previous research [5, 6, 7, 8, 9], and if the effect of the visual scene was not strong enough to override the perception of the acoustically contrasting sound.

For example, it would be expected that a person would incorrectly identify Velcro ripping as paper being torn while watching a person tearing paper. However, it would *not* be expected that someone hearing a foghorn would mistakenly identify this sound as a telephone ringing, even if the visual scene displayed a person answering a telephone. Second and third, confidence ratings of the identifications and the veracity ratings of the sounds were expected to be highest for the actual and similar sounds compared to the contrast sounds. Such results would occur if the visual context impacted and biased the perception of the listener [15, 16, 17, 18]. For example, if listeners are swayed by the visual context and use it help identify the actual and similar sounds, their confidence in their identification and perception of realism should be high. However, if the sounds perceptually mismatch with the visual scene, there should be an impact on the confidence and assessment of the overall realism resulting in lower ratings for both, even though the sound may be correctly identified. Finally, the familiarity ratings for the sounds were expected to be correlated with the number of correct identifications since actual experience with sounds should assist the ability to label them.

2. METHOD

2.1 Participants

There were 45 undergraduate students (31 female and 14 male) who participated in the study for extra credit for psychology courses. The mean age was 20.71 years, and the range was 18 to 22 years and the majority (95%) of them were Caucasian. All participants reported normal hearing and either normal or corrected-to-normal vision. Thirty-five participants completed the sound/video condition; 10 completed the sounds-only control condition.

2.2 Apparatus

The scenes were filmed using a Canon GL1 digital video camcorder. An Audio-technica MB 4000C microphone was used to record the auditory stimuli that were recorded by the experimenters. The video and audio recordings were edited using FinalCut Pro 4.0. The final videos were presented to participants using PowerPoint on Apple Powerbooks with Sony MDR-CD850 stereo headphones.

2.3 Auditory and Visual Stimuli

There were 36 everyday sounds made by objects chosen for use in the experiment (see Table 1) based on data from a previous study [8]. Thirteen of the sounds were the sounds made by the objects in the videos (actual sounds); 10 of the sounds were acoustically similar sounds to those made by the objects in the videos that had been misidentified as the actual sounds (similar sounds); and 13 of the sounds were acoustically dissimilar and had not been confused with the sounds in the videos (contrast sounds). Three of the actual sounds (book closing, stapler stapling, and paper ripping) were also used as similar sounds for the videos.

Table 1: Sound stimuli and their relationship with the videotaped scenes

Actual	Similar	Contrast
3-Ring Binder (closed)	Purse (snapped shut)	Hair Dryer (turned on)
Book (shut)	Balloon (popped)	Vinyl Record (scratched)
Soda Can (crushed)	Book (shut)	Vacuum Cleaner (turned on)
Soda Can (opened)	Stapler (stapling)	Touch-tone Phone (dialed)
Chalkboard (erased)	Eraser (erasing in paper)	Rattle (shaken)
Keys (jingled)	Chains (clinked)	Chalkboard (written on)
Hammer (pounding)	Basketball (bounced)	Tires (Screeching)
Paper (ripping)	Tape (pulled off roll)	Sword (taken out of sheath)
Telephone (ringing)	Alarm Clock (ringing)	Foghorn (blown)
Scissors (snipped)	Whip (cracked)	Baseball (hit with bat)
Spoon (dropped)	Nails (dropped)	Ratchet (turned)
Stapler (stapling)	Cigarette Lighter (flicked)	Glass (breaking)
Velcro (pulled apart)	Paper (ripping)	Saw (sawing wood)

The scenes were the action on the object that produced the actual sound and were staged with a single person in a context where such an action might normally happen. They were videotaped with the target action and sound repeated 3 times. During the recording, the audio was also recorded so that there were other minor ambient sounds available in the soundtrack. After recording was completed, the videos were edited and the sounds were synchronized with the actions for all three types of sounds. The resulting 39 videos were distributed across 3 sets of 13 videos so only one of the scenes was represented in each set, and the sound conditions (actual, similar, and contrast) were counterbalanced across the sets. Due to the small number of trials per individual, the conditions were unequally distributed for each set, such that there were no fewer than 3 and no more than 6 from each condition. This was done to prevent participant bias based on expectations for answers on given trials. Each of the 3 sets of videos were placed in PowerPoint slides in 2 random orders resulting in 6 sets of PowerPoint slides for the sound/video condition procedure.

Two random orders of all 36 sounds were produced and placed in PowerPoint slides for presentation to the participants in the sounds-only condition. The slide used to designate each sound had the number displayed in the middle of the screen that corresponded to the trial on the response sheet.

Finally, 2 random orders of a written list of all 36 sounds were produced that were used for participants in both the sound/video condition and the sounds-only control for the familiarity ratings.

2.4 Procedure

For the sound/video condition participants were randomly assigned to one of the six sets of PowerPoint slides. The experimenter read a set of instructions to the participants while they read along. Participants were told that they would be viewing videotapes of people in 13 everyday situations. They were also told that after each scene, they would be asked to identity the sound the object made and to rate their confidence in their identification (1, not confident, to 7, very confident) and the veracity of the sound (1, not realistic, to 7, very realistic).

Participants completed a practice trial and were allowed to ask questions about the procedure. After they completed the 13 video trials, they were given a written list of all 36 sounds and rated each of them on familiarity (1, not familiar, to 7, very familiar). To finish the procedure, participants completed a brief follow-up questionnaire after which they were fully debriefed.

For the sounds-only condition, participants were randomly assigned to one of the two orders of the 36 sounds. After each sound trial, they made an identification of the sound and rated their confidence in this identification as well as a rating of the sound's veracity. After these trials were completed, they rated the written list of sounds for familiarity. These tasks were the same as those performed by the sound/video condition group, except that this group was not exposed to the videos.

3. RESULTS & DISCUSSION

For the sound/video condition there was one within-subjects independent variable, the sound and video pairings, which had three conditions, actual, similar, and contrast. The dependent variables were the number of correct sound identifications, the ratings of confidence for the sound identifications on a 7-point scale (1, not confident, to 7, very confident), the ratings of veracity of the sound (1, not realistic, to 7, very realistic), and the rating of the familiarity of each sound (1, not familiar, to 7, very familiar). The sounds-only control condition had data for all three dependent variables.[1]

3.1. Sound identifications

For the number of correct identifications for the sounds, a repeated measures ANOVA and follow-up analytical comparisons revealed that the actual sounds (M = 4.06, SD = 1.04) had the highest mean number of correct identifications, followed by the contrast sounds (M = 2.56, SD = 1.05) with the similar sounds (M = .62, SD = .82) showing the lowest mean number of correct identifications, $F(2,66) = 93.31$, $p < .001$, η^2_p =.74. These results provide partial support for the hypothesis since it was expected that the actual and contrasting sounds would have higher identification rates than the similar sounds. Contrary to the hypothesis, it was found that the actual sounds had a higher identification rate than the contrasting sounds. Considering these data as percentages clearly shows the difference in identification rates with actual sounds identified 95%, contrast sounds 61%, and similar sounds 14% of the cases (see Figure 1). Further examination of the incorrect identifications of the similar sounds showed they were misidentified 60% of the time as the sound made by the object in the video; however, the contrast sounds were never identified in this matter. The control group, who only heard the sounds, had an identification rate of only 49%.

[1] The control group for this design was not included in the statistical analyses with the experimental groups due to the different number of stimuli in the control versus experimental conditions. However, the control group data were included in the results to give an *indication* of how people perform these auditory tasks when they have no contextual visual information.

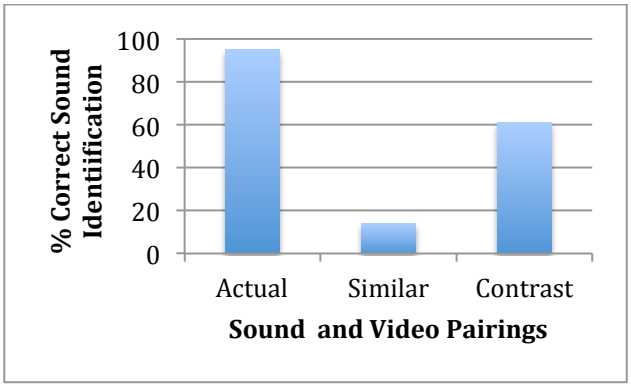


Figure 1: Percentage of correct sound identifications for the sound and video pairings.

3.2 Confidence ratings for sound identifications

For the confidence ratings for the identifications, a repeated measures ANOVA with post hoc analytical comparisons revealed that actual sounds (M = 6.43, SD = .60) showed the highest ratings while there was no difference between the similar (M = 4.64, SD = 1.19) and the contrast (M = 5.00, SD = 1.31) sound ratings, $F(2,66) = 30.11$, $p < .001$, η^2_p = .48 (see Figure 2). These results show partial support for the hypothesis since the actual sounds were given higher confidence ratings than the contrast sounds as predicted, but contrary to the hypothesis, the similar sounds were not rated higher than the contrast sounds and had lower ratings than the actual sounds. The control group's confidence ratings for all sounds showed a base rate that fell within the means of the experimental conditions (M = 5.13, SD = 1.03).

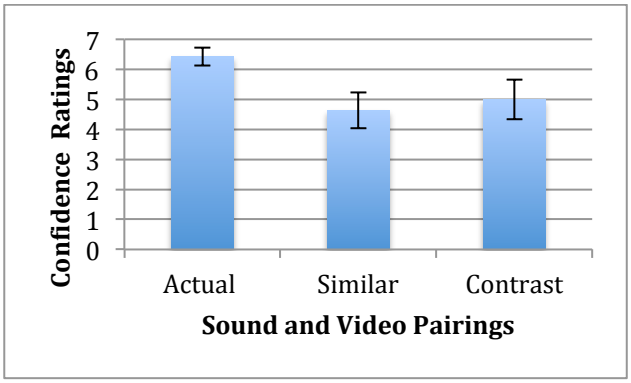


Figure 2: Confidence ratings for sound identifications for the sound and video pairings.

3.3 Veracity ratings for sounds

For the veracity ratings, a repeated measures ANOVA with post hoc analytical comparisons showed that actual sounds were viewed as most realistic (M = 6.39, SD = .54), followed by similar sounds (M = 3.86, SD = 1.13) with contrast sounds having the lowest veracity rating (M = 2.26, SD = 1.65), $F(2,66)$ = 116.50, $p < .001$, η^2_p =.78 (see Figure 3). These results provided partial support for the hypothesis since it was expected that the actual and similar sounds would have higher veracity ratings than the contrast sounds, but it was not

expected the similar sounds would be perceived as less realistic than the actual sounds. The means for the control group indicate the rated realism of sounds only was closest to the actual sound condition ($M = 5.96$, $SD = 1.24$).

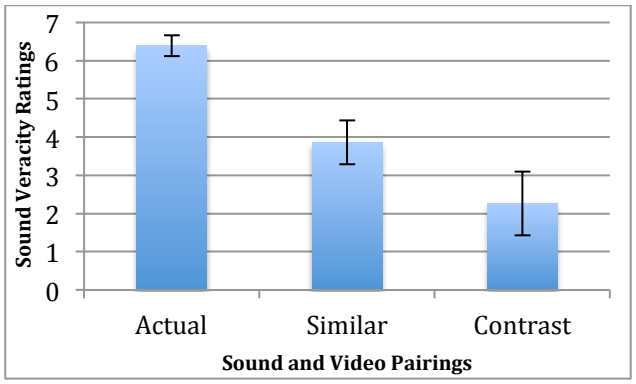


Figure 3: Sound veracity ratings for the sound and video pairings.

3.4 Familiarity ratings for sounds

Finally, ratings of familiarity for the sounds showed that the more familiar the sound was, the higher the number of correct identifications for the actual sounds and for the sounds in the control condition, $r(43) = .44$, $p < .001$. However, the familiarity ratings for the contrast, $r(33) = .23$, $p > .05$ and similar sounds, $r(33) = .17$, $p > 05$, provided no predictive value.

4. CONCLUSION

The results from this study clearly show that people watching videos of actions in which objects are "sounded" impact their perception of the sound. When the sound is the actual sound made or is an acoustically contrasting sound, their ability to make correct identifications is much better than when the sound is acoustically similar. These results even suggest that there is a facilitative effect for seeing the action and hearing the sound at the same time rather than just hearing the sound alone. The inaccurate identifications of the similar sounds show what would be expected from the Foley representations of sounds – people accept the sound as that portrayed by the video. However, it is important to note that in contrast to expectations that similar sounds would be *completely* perceived as real, listeners' confidence in such identifications and their assessment of the realistic nature of the sounds show that they do indeed recognize that the sound is not quite right. Since the stimuli in this study have only one sound that was actively portrayed, it is reasonable to predict that adding more background sound effects and more visual action would lead to people not noticing the discrepancy between the visual scene and an accompanying acoustically similar sound that is not the actual sound made by the object. In such cases, the coconuts banged together would indeed be perceived as horse hoofs galloping across the prairie.

5. ACKNOWLEDGMENT

I would like to thank Tanja Gazibara, Philip Schuman, Natalie Piltz, and J. Allen Lynch for their assistance with this project.

6. REFERENCES

[1] H. Mantell (Ed.) The complete guide to the creation and use of sound effect for film, T.V. and dramatic productions: And for exercising the mind, the ear, the imagination and the pen. Princeton: Films for the Humanities, Inc., 1983.

[2] S. Namba, Y. Hayashi, S. and Wako, "On the synchrony between figure movement and sound change," *Empirical Studies of the Arts*, vol. 21, 177-184, 1998.

[3] E. H. A. Langendijk and A. W. Bronkhorst, "Fidelity of three-dimensional-sound reproduction using a virtual auditory display," *J. Acoustical Soc. Am.,* vol. 107, 582-527, 2000.

[4] N. F. Dixon and L. Spitz, "The detection of auditory visual desynchrony," *Perception*, vol. 9, 719-721, 1980.

[5] J. A. Ballas, "Common factors in the identification of an assortment of brief everyday sounds," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 19, 250-267, 1993.

[6] W. H. Warren, Jr. and R. R. Verbrugge, "Auditory perception of breaking and bouncing events: A case study in ecological acoustics, *Journal of Experimental Psychology: Human Perception and Performance*, vol. 10, 704-712, 1984.

[7] N. J. Lass, S. K. Eastman, W. C. Parrish, K. A. Scherbick, D. M. Ralph, "Listeners' identification of environmental sounds, Perceptual and Motor Skills, vol. 55, 75-78.

[8] T. L. Bonebright, "Perceptual structure of everyday sounds: A multidimensional scaling approach," *Proceedings of 2001 ICAD.*

[9] B. Gygi, G. R. Kidd, and C. S. Watson, "Similarity and categorization of environmental sounds," *Perception and Psychophysics*, vol. 69, 839-855, 2007.

[10] J. A. Ballas and T. Mullins, "Effects of context on the identification of everyday sounds," *Human Perception*, vol. 5, 199-219, 1993.

[11] B. Gygi and V. Shafiro, "The incongruency advantage for environmental sounds presented in natural auditory scenes," vol. 37, 551-565, 2011.

[12] S. J. Boyce and A. Pollatsek, "Identification of objects in scenes: The role of scene background object naming," *Journal of Experimental Psychology: Learning, Memory and Cognition*, vol. 18, 531-543, 1992.

[13] I. Biederman, "Perceiving real-world scenes," *Science*, vol. 177, 77-80, 1972.

[14] J. Davenport and M. C. Potter, "Scene consistency in object and background perception," *Psychological Science*, vol. 15, 559-564, 2004.

[15] H. McGurk and J. W. MacDonald, "Hearing lips and seeing voices, *Nature*, vol. 264, 746-748, 1976.

[16] K. P. Green and A. Gerdeman, "Cross-modal discrepencies in coarticulation and the integration of speech information: The McGurk effect with mismatched vowels," *Journal of Experimental Psychology*, vol. 21, 1409-1426, 1995.

[17] J. Vroomen and B. de Gelder, "Sound enhances visual perception: Cross-modal effects of auditory organization on vision, *Journal of Experimental Psychology*, vol. 26, 1583-1590, 2000.

[18] J. Vroomen and B. de Gelder, "Perceptual effects of cross-modal stimulation: Ventriloquism and the freezing phenomenon," In G. Calvert, C. Spence, and B. E. Stein (Eds.), *Handbook of Multisensory Processes*, pp. 141-152, 2004.

CORRELATIONS AND SCATTERPLOTS : A COMPARISON OF AUDITORY AND VISUAL MODES OF LEARNING AND TESTING

Michael A. Nees

Lafayette College
Department of Psychology
Oechsle Hall, Easton, PA, 18042
neesm@lafayette.edu

ABSTRACT

An experiment examined brief, computer-based modules for teaching and conducting achievement testing of introductory concepts related to correlations and scatterplots. Participants experienced either auditory or visual learning modules followed by either auditory or visual tests of the concepts presented in the modules. Visual modules and tests used on-screen text and visual scatterplots, whereas auditory learning modules and tests presented the same content with text-to-speech (TTS) presentations and auditory versions of scatterplots. Across learning and testing manipulations, no differences were found in the accuracy of responses on the tests, but both auditory learning and auditory testing resulted in longer response times. Results are discussed in the context of computer-based learning and auditory learning and testing as an accommodation.

1. INTRODUCTION

The auditory presentation of text via digital TTS has become more practical to implement in an array of devices, and interfaces that use TTS (e.g., the Siri digital assitant on Apple's iPhone) seem to be growing in popularity. The use of TTS for educational and learning activities also has become more feasible in computers and a multitude of digital devices. In a recent survey, over one third of e-reader users and over half of e-book users rated TTS functionality as "valuable" or "very valuable" [1], and at least some online TTS services (e.g., www.ispeech.org) explicitly tout the value of TTS for translating educational materials to the auditory modality.

TTS technologies can use sound to display text to visually-impaired learners and also to sighted learners with alternative learning preferences. Though computer-based and digital learning technologies have become ubiquitous, research to date largely has not addressed fundamental best-practice questions surrounding the implementation of these technologies for teaching and learning [2]. The redundant presentation of auditory and visual learning materials can be beneficial [3], but the equivalence of auditory-only versus visual-only presentations of materials for pedagogical purposes remains unclear. Similarly, very little research exists on the design of auditory-only tests.

The lack of research comparing the efficacy of visual and auditory presentations of learning and testing materials is problematic for several reasons. First, there seems to be a pervasive assumption that the auditory delivery of text—even very complex text associated with many learning activities—provides learning opportunities that are equivalent to the opportunities offered when the same text is presented visually. This assumption may be flawed when it is applied to the delivery of complex curricula in science and math education, as the demands placed on working memory by the transient nature of the auditory presentation may present memory difficulties (i.e., extraneous cognitive load, see [4]) not encountered by visual learners of the same text.

In addition to the delivery of curricula through sound, a common accommodation for test-takers with visual impairments (or other disabilities) has been the auditory presentation of test questions [5]. Despite the prevalence of this accommodation in both aptitude and achievement testing, researchers have yet to establish the validity of tests administered under oral accommodations. In current practice, often a human reader will administer oral examinations. This presents obvious complications for standardization of testing conditions, and researchers [6] have suggested that a better approach may be to develop "self-voicing" TTS systems to administer oral versions of tests [7]. The comparability of auditory and visual modes of information presentation should be established to ensure that equitable delivery of curricula and fairness in testing can be accomplished in both modalities.

Another known gap exists in the translation of graphical materials in visual texts into auditory representations [8] for both learning and testing. Geisinger [9] pointed out, "…the use of figures and graphs make tests more difficult and typically may alter the cognitive processes employed—because they must be described verbally to the test taker with visual impairment" (pp. 131). Auditory graphs offer a promising alternative to verbal descriptions for translating graphs into sound, as emergent percepts of data patterns may function similarly in auditory and visual graphs [10]. Research [11] has shown that auditory versions of scatterplots are as effective as visual representations for conveying correlations.

The current study examined auditory and visual learning and testing of introductory statistical concepts about correlation and scatterplots in a sample of university students with no prior formal education in statistics. The use of TTS and auditory versions of scatterplots was compared to visual presentation of text and graphical scatterplots with a 2 (learning module: visual or auditory) X 2 (test format: visual or auditory) X 2 (question type: scatterplot or no scatterplot) mixed design. The study was designed to examine: 1) the efficacy of both auditory and visual learning; 2) the comparability of auditory and visual testing; 3) the possibility of interactions between modes of learning and testing; and 4) the possibility of

differential effects for test questions that required or did not require judgments about (auditory or visual) scatterplots.

2. METHOD

2.1. Participants

Participants (N = 41; 20 females; M age = 20.0, SD = 1.9 years) were recruited from undergraduate psychology courses at the Georgia Institute of Technology. Participants were excluded if they had taken a statistics course at the high school or college level, and all participants reported normal or corrected-to-normal hearing and vision.

2.2. Stimuli

A brief (approximately 3000 word) script of a lesson on correlations and scatterplots was prepared. The lesson covered basic concepts such as the direction and strength of correlations, interpreting r values, and reading bivariate scatterplots of data. This lesson was the basis of the respective auditory and visual learning modules, both of which were approximately 20 min in duration.

For the visual module, the text was presented on the computer screen in complete sentences (from one up to several sentences at once) at a pace that was controlled by the computer. The duration of the text presentations was yoked to the duration of the corresponding TTS audio file from the auditory module (described below), thus the learning modules were exactly matched in duration. Visual examples of scatterplots used in the module were made using Microsoft Excel and were displayed alone on the screen for 5 s. To ensure that the information contained in the visual scatterplots was commensurate with that of the auditory scatterplots, all visual scatterplots were stripped of axis labels; only data points showing the relationship between the two variables depicted, which were described in the text, were displayed.

For the auditory learning module, TTS conversions of the text of the visual module were made with the demo function of the TTS engine at http://www.ispeech .org in early 2011[1]. TTS was created using the "English male" voice (now "US English male") at the "normal" (i.e., default) speed setting. The text of the visual module was converted to mp3 files from the website. Exact text from the visual modules was entered into the TTS engine with two exceptions: 1) where appropriate, the text was modified to reflect the auditory nature of the module (e.g., "scatterplot" was changed to "auditory scatterplot"); and 2) numbers and symbols were entered into the TTS engine as words to ensure that the auditory speech was intelligible for all text elements from the visual module (e.g., "r =" was voiced as "*are equals*"). The data from each scatterplot in the visual module were sonified into auditory graphs using the Sonification Sandbox [12] software. All scatterplots were sonified to be 5 s in duration in the range of notes C4 (MIDI note 60, 262.6 Hz) to C8 (MIDI note 108, 4186.0 Hz) using a positive polarity mapping and the MIDI piano timbre.

The visual test consisted of 20 multiple-choice questions. Test questions were comparable to the types of practice and test questions on correlations found in introductory statistics texts. Each question was displayed on the screen in its entirety with each of the four possible answers visible. The auditory test presented the exact same questions and answers with TTS. Each question was read in its entirety, followed by each of the four possible answers in succession. The test was designed such that half of the questions were conceptual in nature and did not display a scatterplot representation of the data, while half of the questions displayed one or more scatterplots as part of the question or answers. At the beginning of each test, participants were given a brief (one paragraph) overview of either auditory or visual scatterplots (depending on the test format condition). The overview was necessary to explain the respective representations to participants who had experienced the learning module in a different modality from the test format (e.g., participants who experienced the visual learning model needed a brief description of how the auditory scatterplots represented data).

2.3. Procedure

Following informed consent, participants were randomly assigned to one of the four factorial combinations of the 2 (learning: auditory versus visual) x 2 (test: auditory versus visual) between-subjects independent variables. Participants were seated at a computer in front of a 17 in (43.2 cm) Dell LCD computer monitor. A computer program made with Adobe Director presented stimuli and collected data. Auditory stimuli were presented with Sennheiser HD 202 headphones. All participants wore headphones during the study, though no sounds were presented to participants assigned to visual learning conditions and visual testing conditions. Similarly, the computer screen was blank during auditory conditions of the study. Participants experienced either the auditory or visual learning module, followed by either the auditory or visual test. The 20 test questions were presented in a random order for each participant, and both responses and response times were recorded. The response time for a trial was operationally defined as the duration between the onset of the question (i.e., the appearance of the question on the screen for the visual test or the beginning of the TTS audio reading of the question for the auditory test) and the logging of a response to the test question. Participants in either condition could log a response at any time; participants in the auditory test condition were not obligated to listen to the entire question and set of answers. Following the test, participants completed the NASA-TLX [13] measure of subjective workload.

3. RESULTS

Analyses were conducted using mixed 2 (learning module: visual or auditory) X 2 (test format: visual or auditory) X 2 (question type: scatterplot or no scatterplot) ANOVAs on both the number of correct answers on the test and the response times to test questions. For the number of correct answers, the main effects of learning module, $F(1,37) = 0.37$, $p = .55$, and test format, $F(1,37) = 3.26$, $p = .08$, were not statistically

[1] At the time of this submission, the website appeared to have made minor modifications to the interface and TTS algorithms since the stimuli were created.

significant. The effect of question type was significant, $F(1,37) = 13.24$, $p = .001$, $\eta^2_P = .26$; participants correctly answered more questions without scatterplots ($M = 6.37$, $SE = .32$) than with scatterplots ($M = 5.28$, $SE = .24$). There were no statistically significant interactions (p values ranged from .14 to 89). Of note, the test exhibited neither ceiling nor floor effects; chance performance would have resulted in $M = 2.5$ correct answers in each condition. Results are shown in Figure 1.

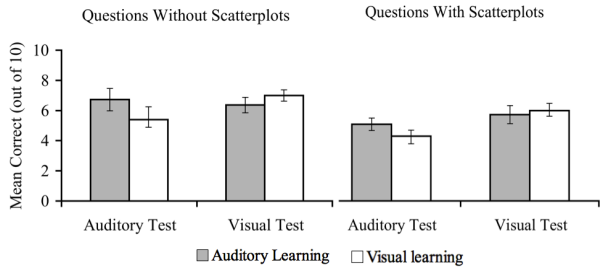


Figure 1: Mean numbers of correct answers across conditions. Error bars represent standard error.

For the response times (reported in s) to test questions, all main effects were significant. For the learning module independent variable, $F(1,37) = 5.23$, $p = .03$, $\eta^2_P = .12$, participants' mean response times were significantly faster if they had learned the material from the visual module ($M = 30.3$, $SE = 1.4$) as compared to the auditory module ($M = 25.8$, $SE = 1.5$). For the test format independent variable, $F(1,37) = 54.37$, $p < .001$, $\eta^2_P = .60$, participants' mean response times were significantly faster if they took the visual version of the test ($M = 20.7$, $SE = 1.4$) as compared to the auditory version ($M = 35.3$, $SE = 1.4$). For the question type independent variable, $F(1,37) = 13.00$, $p < .001$, $\eta^2_P = .26$, participants' mean response times were significantly faster for questions without scatterplots ($M = 26.5$, $SE = 1.1$) as compared to questions with scatterplots ($M = 29.6$, $SE = 1.1$). The interaction of test format with question type was also significant, $F(1,37) = 30.14$, $p < .001$, $\eta^2_P = .45$. The interaction was reflected in the fact that participants taking the auditory version of the test were slower to provide a response to questions with scatterplots ($M = 31.4$, $SE = 1.5$) as compared to questions without scatterplots ($M = 39.3$, $SE = 1.5$), but participants taking the visual version of the test did not show a difference for questions without scatterplots ($M = 21.6$, $SE = 1.5$) as compared to questions with scatterplots ($M = 19.9$, $SE = 1.6$). Results are shown in Figure 2.

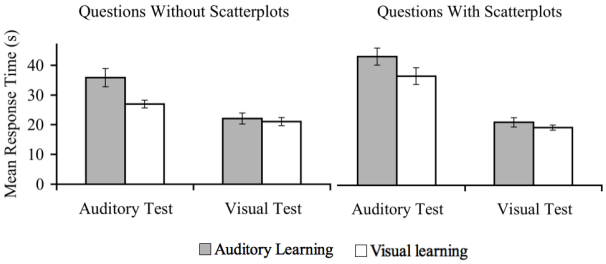


Figure 2: Mean response times across conditions. Error bars represent standard error.

One obvious interpretation of the disparities in response times is that participants in the visual condition simply read the questions and answers faster than the TTS presented the questions and answers in the auditory version of the test. This interpretation would be supported if the difference between mean auditory and visual test response times increased as the duration of the auditory questions and answers increased. To examine this possibility, an exploratory correlation showed that, across the 20 different questions, the duration of the audio version of the question and answers (i.e., the time required for participants to hear the question and all four answers in the auditory test condition) and the difference in mean response times for the visual versus auditory test conditions were not related, $r(19) = .38$, $p = .10$. Though this relationship (see Figure 3) might have reached statistical significance with a larger sample of questions, the pattern of results showed that the duration of the auditory test questions alone did not account for tendency of auditory test-takers to require a longer response time.

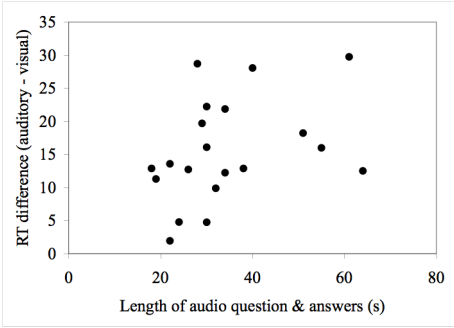


Figure 3: The length of audio questions and answers as a function of the mean response time difference (visual format subtracted from auditory format) for each of the test questions.

Finally, a 2 (learning module: visual or auditory) X 2 (test format: visual or auditory) ANOVA was performed on the NASA-TLX composite scores. The main effects of learning module, $F(1,37) = 0.12$, $p = .73$, and the interaction of learning module with test format, $F(1,37) = 0.89$, $p = .35$, were not statistically significant. The main effect of test format was significant, $F(1,37) = 8.04$, $p = .007$, $\eta^2_P = .18$. Participants in the auditory test condition experience greater perceived workload ($M = 10.73$, $SE = 0.64$) than participants in the visual test condition ($M = 8.14$, $SE = 0.65$). Results are shown in Figure 4.

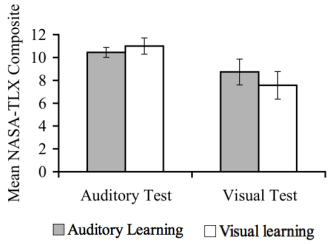


Figure 4: Mean NASA-TLX composite scores. Error bars represent standard error.

4. DISCUSSION

Participants with no prior educational background in statistics learned simple concepts about bivariate correlations equally well—as measured by correctness of responses to test

questions—from visual modules (that used text with traditional visual scatterplots) and auditory modules (that used TTS with auditory scatterplots). Auditory learning modules and auditory versions of achievement tests may represent a viable alternative to visual presentation of materials during learning and testing. Accommodated versions of standardized tests and achievement tests that use oral presentation of questions to date have not found an adequate solution to translating test questions involving graphs and diagrams to the auditory modality. Verbal descriptions of visual figures may be insufficient to offer comparable information, but auditory representations of graphical data may help to fill this considerable gap.

Participants who learned the material with the auditory module took about 5 s longer on average per question on the test than participants who had learned the material with the visual modules, and this effect was present regardless of (i.e., collapsed across) the format of the test. Participants taking auditory versions of the test took about 15 s longer per question on average to register a response to the test question, and responses were even slower with auditory testing for questions that featured an auditory scatterplot as part of the question (as opposed to conceptual questions that featured only spoken words with no scatterplot). The data suggested that longer response times for the auditory version of the test were not simply attributable to the durations of the auditory test questions.

Learners who are assessed with auditory tests may need to be given longer to complete the test. Extended time during testing is another common accommodation that is often implemented in conjunction with auditory presentation of questions. Often, the amount of extra time given seems to be arbitrarily chosen as “time and a half” or “double time.” Studies like this one may be able to offer empirical guidelines for the amount of extra time needed to achieve comparable mean performance across testing formats. The significant difference in perceived workload did not correspond to an objective decrease in test performance as measured by the number of correct answers, but the perceived workload could potentially have detrimental effects on test performance in assessment scenarios that run longer than the brief test here.

5. CONCLUSIONS

TTS auditory versions of learning materials with auditory graphs may offer a comparable alternative to traditional visual learning materials for teaching basic statistics concepts. Perhaps even more importantly, TTS versions of tests with auditory graphs may offer a standardized means of assessing achievement of basic statistics (and perhaps other math) concepts in the auditory modality that is comparable to the visual tests currently used in learning assessment, though the current study’s results suggest that TTS test-takers may require more time to complete assessments. The finding that auditory test-takers perceived higher subjective workload warrants further investigation.

6. ACKNOWLEDGMENT

This research was conducted at the Georgia Tech Sonification Lab with support from Bruce Walker. Zia Drakhshandeh created some stimuli for this experiment, and Michelle Han collected data. Their contributions are gratefully acknowledged.

7. REFERENCES

[1] N. Foasberg, "Adoption of e-book readers among college students: A survey," *Information Technology and Libraries,* vol. September, pp. 108-128, 2011.

[2] D. A. Cook, "The research we still are not doing: An agenda for the study of computer-based learning," *Academic Medicine,* vol. 80, pp. 541-548, 2005.

[3] R. Moreno and R. E. Mayer, "Verbal redundancy in multimedia learning: When reading helps listening," *Journal of Educational Psychology,* vol. 94, pp. 156-163, 2002.

[4] J. Sweller, "Cognitive load theory, learning difficulty, and instructional design," *Learning and Instruction,* vol. 4, pp. 295-312, 1994.

[5] S. G. Sireci, S. E. Scarpati, and S. Li, "Test accommodations for students with disabilities: An analysis of the interaction hypothesis.," *Review of Educational Research,* pp. 457-490, 2005.

[6] E. G. Hansen, M. J. Lee, and D. C. Forer, "A "self-voicing" test for individuals with visual impairment," *Journal of Visual Impairment & Blindness,* pp. 273-275, 2002.

[7] R. P. Dolan, T. E. Hall, M. Banerjee, E. Chun, and N. Strangman, "Applying principles of universal design to test delivery: The effect of computer-based read-aloud on test performance of high school students with learning disabilities.," *Journal of Technology, Learning, and Assessment,* vol. 3, 2005.

[8] M. A. Nees and B. N. Walker, "Data density and trend reversals in auditory graphs: Effects on point estimation and trend identification tasks," *ACM Transactions on Applied Perception,* vol. 5, Article 13, 2008.

[9] K. F. Geisinger, "Psychometric issues in testing students with disabilities," *Applied Measurement in Education,* vol. 7, pp. 121-140, 1994.

[10] M. A. Nees and B. N. Walker, "Listener, task, and auditory graph: Toward a conceptual model of auditory graph comprehension," in *International Conference on Auditory Display (ICAD2007)*, Montreal, Canada, 2007, pp. 266-273.

[11] J. H. Flowers, D. C. Buhman, and K. D. Turnage, "Cross-modal equivalence of visual and auditory scatterplots for exploring bivariate data samples.," *Human Factors,* vol. 39, pp. 341-351, 1997.

[12] B. K. Davison and B. N. Walker, "Sonification Sandbox reconstruction: Software standard for auditory graphs," in *ICAD 07 - Thirteenth Annual Conference on Auditory Display*, Montreal, Canada (26-29 June), 2007, p. TBD.

[13] S. G. Hart and L. E. Staveland, "Development of the NASA-TLX (Task Load Index): Results of empirical and theoretical research," in *Human Mental Workload*, P. A. Hancock and N. Meshkati, Eds. Amsterdam: North Holland Press, 1988, pp. 239-250.

SONIFICATION AS A SOCIAL RIGHT IMPLEMENTATION

Pablo Revuelta Sanz

Carlos III University of Madrid,
Electronic Technology department,
Av. Gregorio Peces Barba 1, 28918, Spain
prevuelt@ing.uc3m.es

Belén Ruiz Mezcua

Carlos III University of Madrid,
Computer Science department,
Av. Gregorio Peces Barba 1, 28918, Spain
bruiz@inf.uc3m.es

José M. Sánchez Pena

Carlos III University of Madrid,
Electronic Technology department,
Av. Gregorio Peces Barba 1, 28918, Spain
jmpena@ing.uc3m.es

ABSTRACT

The rights discussion begins in Europe, with Modernity, in the 17th century. In this historical moment, social and equal rights are supposed to be universal, and are used to fight against absolutism and the religion hierarchy. However, this egalitarian paradigm has not been applied in such a radical way, allowing some extra-rights environments, which keep working with the *ancient régime* way of life. In the present world, we can identify many people who cannot behave as others do, because of some unwanted circumstances, which diminish their capabilities. We can talk, in these cases, about unimplemented rights. In this paper we discuss whether solutions to disabilities and, more specifically, some applications of sonification, can be treated as a right's implementation and when. We also discuss the limits of the rights under an economical system such as capitalism, and what kind of solutions should be found.

1. INTRODUCTION

The rights discussion starts, as we know it nowadays, in the western world, specifically in Europe, during the so called Enlightens, or 17th century. The Modern way of thinking puts into question the absolute power of kings and religion in the late European middle-age by means of this term.

Although we can search for the first philosophical discussions in the ancient Greece, with Plato, Aristotle, Socrates or Diogenes, we will have to wait to Spinoza [1], among others, to listen to human based rights vindications. This new proposal is revolutionary regarding the *ancient régime*, theocrat and based on vassalage relations and earth subjection [2]. Since several decades before, bourgeoisie was starting its economical revolution which needed new formal and legal structures, much more flexible and autonomous than the absolutism's ones [3].

The subsequent French Revolution, in 1789, sets the basis for the Modern conception of rights, with the "Universal Rights Declaration of Men and Citizens", enacted in the same year.

1.1. The rights in the Modern culture

The main points of this new culture can be summarized in the following points:

- Every person is son/daughter of God
- Every person is born with identical basic rights
- These rights cannot be sold, bought or transferred
- The role of the state is to ensure them

This school, born mainly in England, but also in France with other thinkers such as Rousseau [4], is known as contractualism or *iusnaturalism*, since they talk about natural rights, inalienable and directly given by God in the natural state. This position, likewise, was supported by other conceptions of ethic; Rationalism and, after that, criticism, with Kant as prime defender, proposes that rights come from rational capacity. This capacity, exclusively human (and maybe also of aliens or God) imposes some specific uses of action capabilities, regarding the so-called *categorical imperative* to those persons (or beings) capable of universalizing the rules of their behavior [5]. The rights are the minimum rational (and coactive) laws or norms that allow every person doing whatever they want, in egalitarian conditions regarding the others. This is called, in political philosophy, the conditions of the negative freedom.

1.2. The contractualism

The name of contractualism comes from the solution to the nature state (supposed to be the original one) given by their proposers. Following Hobbes [6], for example, the way humans achieve to overpass the natural (and violent) situation is performing some agreements (contracts) which make the force

of most of them stronger than the force of each one of their members. The human being evolutes to a normative one, where freedom is sacrificed in the shrine of the security. This new society institutes the Modern State as today we know it.

This study is not the correct place to discuss this position, how liberal anthropology interfered Hobbes analysis of the original situation or how this contract is signed and by whom (see, for example, [7]).

What is important for our analysis is the fact that laws, coming from social contracts, democrat government or from a dictatorship regime, always involve rights, i.e., the capability of doing something and not being punished for that. Likewise, rights involve obligations to another party. If I have the right of living, everyone else has the obligation of respecting my life. As simple as that. But no so simple.

The main discussion is which rights (and, therefore, which obligations) must be sanctioned by the contract, and which should not. The obvious problems of this approach, developed under the liberal paradigm and, more specifically, under its economical implementation, will be discussed in the next section.

1.3. The limits under the capitalism

The rights discussion, as we saw, refers to an egalitarian idea of human societies. However, people act in a different way, and this fact may yield to differences in what they have, or do, in their lives. This is the base of the so-called *meritocracy*.

Mainly heritage, but also other social devices such as favoritisms, racism, sexism, etc. can generate non-egalitarian points of depart for every new person coming into the world, and distort the ideal liberal society producing a classist one.

Moreover, the social contract was not signed by anyone alive today, but all of us are forced to obey it [8]. Thus, the rights became positive and traditional (statutory) instead of being rationally supported.

Finally, there are different kinds of rights, and two former groups among them:

- Those which are material cost free (such as free speech right), and
- Those which are not (right to a dignified living situation, right to work, etc.).

The main problem that the society under capitalism has to face is the deficient material implementation of some rights.

Since capitalism is an auto-regulated economical system, it has its internal rules. These rules, however, do not have anything to do with what we call social and political rights, except one: follow your own interest.

With this constraint, it is hard to understand A. Smith's proposal of the invisible hand [9], and a sight in nowadays world may discourage everyone of thinking in such an innocent way. The tragedy of commons [10] should be the final picot to this school. This philosophy has different results regarding the human rights:

- No planning over the present generation (temporal constraint). Ecology is seen as an enemy of business.
- No planning out of profitable niches (local constraint).

Commodities are only made if the result is a profit, i.e., is the Money-Commodity-Money' wheel turns [3]. Strange illnesses research, environment responsibility, labor improvements, ecological fingerprint and any other *common expense* are seen as a waste and, hence, not taken into account by the capitalist logic by its own.

However, new rights are emerging, apart from capitalism, since this system will never cover some aspects which, as it will be discussed in the following section, may be treated as rights. Among them, we will focus in this work on a specific one: the right of the blind people to access public visual information.

2. SHOULD SONIFICATION BE A RIGHT?

Being born blind, or becoming blind by any cause, eliminates a part of the perception capabilities. The same occurs, in different ways, with other disabilities. This constraint makes it difficult to perform some common life tasks, which are taken as basic rights in most of the Constitutions, such as movement, working or access to information, among others. Before 2006, when the United Nations signed the Convention of the Rights of Persons with Disabilities [11], some other essays had been proposed to address this problem: the Declaration on the Rights of Mentally Retarded Persons [12] or the Declaration on the Rights of Disabled Persons [13].

Thus, disabilities which restrain some capabilities in some social environments should be read as rights diminution.

There is another way to support this relation. J. Rawls [14] proposed the *veil of ignorance*, to imagine the situation where you do not know your identity, gender, race, social class and, we could add, disability. In such situation, you are asked to decide how your society should work, regarding rights and obligations. The answer to that question, given the veil of ignorance, shows us if we consider something as a right.

2.1. Where and how are auditory displays already taken as rights implementations

In fact, accessibility is already seen as a rights matter in many countries, which have developed a new legal corpus to minimize the social, material and psychological effects of the different disabilities (see, for example, [15]).

In this work, we will only focus on visual information accessibility through sonification. Other ways of providing accessible information for the blind persons will not be discussed in this work.

Sonification should only be treated as a human right implementation when it minimizes the effect of a disability regarding some right enjoyment. This has been the goal of some proposed sonification devices, since the end of the XIX century [16]. Many other assistive products based on sonification in this line have been proposed (see [17] for a review).

We can find laws, regulations and initiatives in the following environments, implementing sonification as rights and not only as services: TV, cinema and other audiovisual spectacles [18], museums [19], public transport [20] or education [21].

2.2. Who should be obliged by this right

Each time we recognize a right, a correlative obligation is automatically generated. In other words, no right is given for free.

In the case of providing accessibility to visual information for the blinds (as some sonification projects do), there is, likewise, an economic cost. The answer to whom should pay that cost, inside the liberal paradigm, would be the user, who is, at the end, the final responsible of his/her chance.

However, in this point, we should not talk about rights, but about business. Rights, obviously, cannot be sold or bought.

Some other institutions have been proposed to solve some rights disruptions in special cases, such as NGO's during humanitarian actions. These organisms, depending on the charity of their supporters, can never guarantee a right's implementation. The precariousness will threaten every single day of existence of the right under these conditions.

Finally, a social consensus to recognize something as a right is the only way to convert this proposal into a material right. Likewise, the cost should be, then, assumed by every single person who has supported this right constitution.

3. CONCLUSIONS

Sonification, when it tries to overcome visual limitations due to different disabilities or circumstances, can be treated as a right. However, this point of view must surpass the narrow liberal paradigm regarding material rights.

Likewise, rights impose obligations to a second party, which should assume the economical cost of the audiovisual accessibility. If these costs are not assumed, the blinds will depend on the charity or on their own savings.

Sonification is, essentially, a good candidate to implement new and uprising rights.

4. REFERENCES

[1] B. Spinoza, *Ethics*, Tredition, 2011.

[2] P. Anderson, *Lineages of the Absolutist State*, Verso, 1996.

[3] K. Marx, *Capital: Volume 1: A Critique of Political Economy*, Penguin Classics, 1992.

[4] J.-J. Rousseau, *The Social Contract*, CreateSpace, 2012.

[5] I. Kant, *Fundamental Principles of the Metaphysic of Morals*, CreateSpace, 2011.

[6] T. Hobbes, *Leviathan*, Empire Books, 2011.

[7] F. Engels, *The Origin of the Family, Private Property and the State*, Penguin Classics; Revised edition, 2010.

[8] J. Wolff, *An Introduction to Political Philosophy*, Oxford University Press, USA; Revised edition, 2006.

[9] A. Smith, *The Wealth of Nations*, Simon & Brown, 2012.

[10] T. R. Machan, *The Commons: Its Tragedies and Other Follies*, Hoover Institution Press; 1st edition, 2001.

[11] UN, *Convention of the Rights of Persons with Disabilities*, http://www.un.org/disabilities/documents/convention/convoptprot-e.pdf, 2006.

[12] UN, *Declaration on the Rights of Mentally Retarded Persons*, http://www2.ohchr.org/english/law/res2856.htm, 1971.

[13] UN, *Declaration on the Rights of Disabled Persons*, http://www2.ohchr.org/english/law/res3447.htm, 1975.

[14] J. Rawls, *A Theory of Justice,* Original Edition, Belknap Press of Harvard University Press, 2005.

[15] Dept. of Justice, ADA *A Guide to Disability Rights Laws*, http://www.ada.gov/cguide.htm, 2012.

[16] W. Starkiewicz and T. Kuliszewski, "The 80-channel elektroftalm." *Proceedings of the lnternational Congress Technology Blindness, Am.Found.Blindness* . New York, 1963.

[17] Revuelta Sanz, P., Ruiz Mezcua, B., & Sánchez Pena, J. M., *ICTs for Orientation and Mobility for Blind People. A State of the Art,* In: ICTs for Healthcare and Social Services: Developments and Applications, I. Maria Miranda & M. Manuela Cruz-Cunha (eds), 2011.

[18] FCC, Federal Communications Commission. *Closed captioning of video programming*, 2009, from http://www.fcc.gov/cgb/dro/caption.html.

[19] AENOR. Norma UNE 153020. *Audiodescripción para personas con discapacidad visual. Requisitos para la audiodescripción y elaboración de audioguías*. 2005.

[20] The Guide Dogs for the Blind Association, *Audio-visual information systems on Buses A Joint Statement*, 2012.

[21] Texas School for the Blind and Visually Impaired, *National Agenda for the Education of Children and Youths with Vision Impairments, including Multiple Disabilities*, AFB Press, 1995.

PHYSICAL NAVIGATION OF VIRTUAL TIMBRE SPACES WITH TIMBREID AND DILIB

William Brent

American University
Audio Technology Program
4400 Massachusetts Ave NW
Washington DC

ABSTRACT

This paper summarizes recent development of two open source software libraries that enable auditory display in Pure Data (Pd), and describes developing projects that were achieved using the two packages in tandem. The timbreID feature extraction and classification library enables real- and non-real-time audio analysis via high-level modules that can be programmed for a variety of purposes. DILib (the Digital Instrument Library) provides software tools for accessing and managing gestural control streams as captured by inexpensive, widely available sensor hardware. Realized at the intersection of these software packages, three applications are discussed from technological and performative viewpoints: a system for navigating visual timbre spaces with gestures drawn from full body tracking, a similar system based on open-air infrared fingertip tracking, and the *Gesturally Extended Piano*—an augmented instrument controller that uses piano performance gestures to create visually explicit action-sound relationships.

1. INTRODUCTION

Among available music information retrieval software packages (e.g., [1][2][3][4]), those designed for use in real-time multimedia programming environments are especially valuable for visually-based audio browsing and the performance of live computer music. Such software allows artists to analyze, organize, and reshape immense collections of digitally stored sound with sophistication and relative ease. Parallel to this development—as interest in embodied computer music practices continues to grow—tools that enable the high-speed capture of body movement data have also reached a high level of refinement.

This paper begins by summarizing recent development of two software libraries for Pure Data (Pd) [5], a popular open source multimedia programming environment. The timbreID audio feature extraction and classification library enables real- and non-real-time audio analysis via high-level modules that can be programmed for use in a variety of contexts. Provided example applications include real-time speech recognition, instrument identification, target-based granular synthesis, and various types of sound visualization. The Digital Instrument Library (DILib) provides software tools for accessing and managing gesturally-oriented control streams as captured by increasingly sophisticated yet inexpensive sensor hardware. These include accelerometers, multi-touch surfaces, body tracking systems, and high frame rate digital cameras that can be used for a number of computer vision strategies.

The concerns of these two projects are distinct, but a spectrum of applications exists at their intersection that encompasses purely research-oriented sound exploration tools as well as full-fledged musical instruments. Use of physical gesture information beyond that offered by standard computer input devices enhances applications along this spectrum considerably, making it possible to achieve customized multi-modal relationships with audio based on sound, sight, and touch. The final section of this paper describes three developing projects that explore these possibilities in Pd using timbreID and DILib. Both libraries have been released under the GNU GPL as open source projects, with the intention of promoting novel modes of sound exploration and digital music performance based on freely designed action-sound relationships.

2. AUDITORY DISPLAY WITH TIMBREID

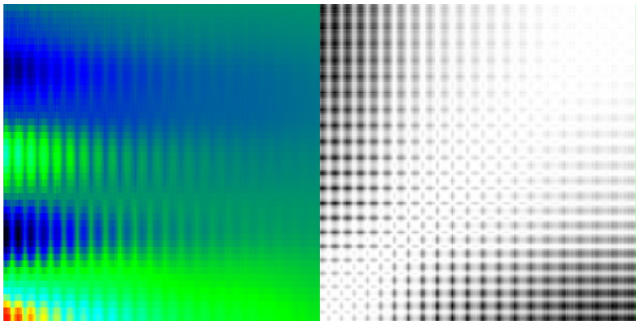


Figure 1: A Bark-frequency cepstrogram (left) and similarity matrix (right) of a tam-tam strike.

Originally described in [6], components of the timbreID library can be used for many different purposes. The current release features improvements and additions to the core analysis and data management objects as well as to the accompanying examples package. Here, we will only summarize example applications that are directly useful for auditory display, with the most significant items being spectrogram, cepstrogram, and similarity matrix plotting tools, and improved functionality of the timbre space plotter. Figure 1 shows a Bark-frequency cepstrogram and similarity matrix of a tam-tam strike that were generated using these tools. Mel- and Bark-frequency cepstrum remain popular as compact descriptors of timbre, but the choice of an optimal range of coefficients for identification tasks requires judgment based on the particular sounds and circumstances. Visualization of cepstral information in the form of a cepstrogram is useful for understanding how individual coefficients vary over the course of specific sounds, and can be a valuable aid in making these kinds of choices.

Plotting segments of audio in relation to their quantifiable features is another technique for understanding relationships between sounds, as well as for designing large and small scale sound sequences based on timbre. In this type of plot, points can be made to represent audio segments of a fixed grain size or entire sound events, and can be auditioned by moving a cursor within range. Figure 2 shows this tool as realized using timbreID, with a collection of piano samples as the objects of analysis. Grains of audio are spaced along the horizontal and vertical axes according to amplitude and spectral centroid, respectively. The axes of the plot can be chosen based on available audio features, and all feature data is displayed for the most recently browsed grain in the information panel shown on the left. Individual audio features can also be plotted against time to reveal dimensions of timbre relative to small and large scale temporal structure.

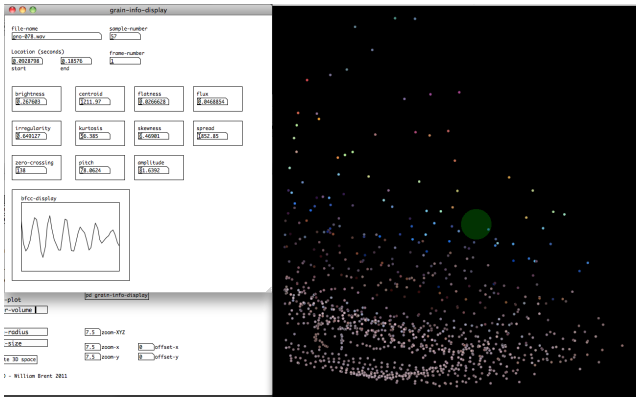


Figure 2: Timbre space plot of piano audio grains.

The main limitation of this basic plotting system is connected with dimensionality. Plots can be viewed and rotated in three dimensions, but only navigated and auditioned in two dimensions with a standard computer mouse. Considering the multi-dimensional nature of timbre, this is a very significant shortcoming. One alternative is to navigate the space based on the qualities of sounds captured by a microphone in real-time. By harvesting the first three Bark-frequency cepstral coefficients (BFCCs) of a live signal as it changes over time, the input sound can be used as a type of cursor moving in three dimensional space. Additional BFCCs can be used to further increase dimensionality, but attempting to make changes in any one dimension by altering the timbre of the input sound does not result in a high degree of control. Further, even in three dimensions, the process of navigating in this manner is very difficult to conceptualize visually. For better results, we need access to control streams from gesture input systems more sophisticated than the standard computer mouse.

3. GESTURE ACQUISITION WITH DILIB

The experience of using interactive sound visualization systems changes fundamentally when different types of body movement are introduced as sources of control. Research in the field of Human Computer Interaction (HCI) has yielded many robust options for capturing physical movement information with minimal encumbrance. The associated hardware and software are increasingly accessible for use within flexible environments like Pure Data, a situation that has encouraged widespread artistic application of these techniques. Moving beyond basic access, a Pure Data library is needed for parsing/routing data streams and generating additional higher level features based on raw tracking information. DILib (originally presented in [7]) aims to meet this demand.

DILib accounts for many different sources of gestural control data. Most relevant to the discussion here are those based on infrared (IR) blob tracking and full body tracking. IR blob tracking has been used as a reliable means of capturing motion information in a variety of contexts. The basic method is to shine a particular wavelength of IR light on a scene, and place highly reflective markers on key points of a moving body. Near the light source, a camera fitted with a bandpass filter tuned to the same IR wavelength observes the scene. Frames in the digital video stream are then subjected to some basic pre-processing before being fed to a blob tracking algorithm. After these steps, objects reflecting a relatively high amount of IR light back to the camera will appear in the video stream as white blobs, while less reflective objects are rendered completely black. Thus, motion within a diverse scene can be reduced to just a few key points of interest.

A significant problem associated with this technique has to do with distinguishing between the tracked blobs. To overcome this, some type of history and analysis of the blob trajectories must be maintained in software. DILib's IR blob tracking module was built using objects in the Graphics Environment for Multimedia [8], and core DILib objects for managing blob continuity and extracting higher level features from blob position data. These features include distances, angles, and centroids between pairs of points, and delta values of individual points across frames. Specific gestures (e.g., pinching and rotation with the fingertips) can be identified based on these features in order to offer different classes of control over synthesis and spatial navigation.

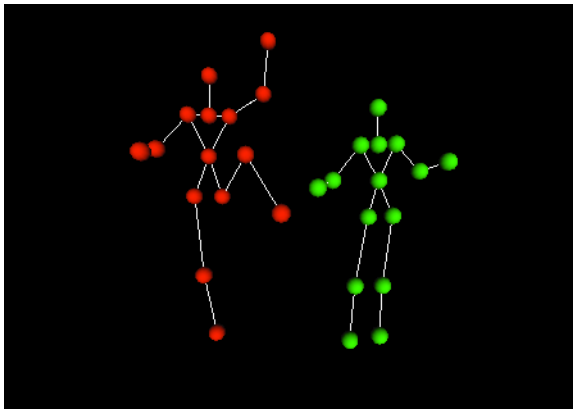


Figure 3: OSCeleton frame data rendered in Pd/GEM with DILib

More sophisticated feature analysis is required for full body tracking, where raw sensor information must be interpreted relative to a model of human movement. DILib's module for body tracking relies entirely on external software for this fundamental step. OSCeleton[1] is open source multi-platform software that interprets data from Microsoft's Kinect sensor and produces three-dimensional coordinates for the primary points of a body being tracked. Its output can be received in Pd via OSC messages, where DILib offers objects for managing data streams of multiple users,

[1] https://github.com/Sensebloom/OSCeleton

graphical rendering of the skeleton frame (shown in Figure 3), and generation of relative data (e.g., distances between extremities, angles at the elbows and knees, etc.).

An important variety of relative data is the offset of an extremity from its attaching joint, such as the three-dimensional position of the right hand in relation to the right shoulder as an origin. Using this approach, the raw coordinate of a user's hand in the entire scene can be polled to control global aspects of a system, while its offset from the shoulder maintains a high degree of independence and is suitable for control over more specific aspects. In this body tracking module and more generally, a central aim of DILib is to facilitate the design of systems that produce complex but consistent consequences in response to changes in basic sources of data. As with acoustic instruments, such systems present an interesting set of constraints, where individual parameters can be modified with near—but not complete—independence.

4. APPLICATIONS

This section reviews characteristics of three real-time sound exploration/instrument systems. In all cases, a fundamental concern is the pursuit of methods for translating continuous movement data into continuous changes in timbre. A second (and somewhat contradictory) function is transformation of this core sound via a layer of dynamically routed signal processing modules.

4.1. Embodied Timbre Space Navigation

In the context of a timbre space, the skeleton frame data described in Section 3 can be used in any number of ways. In the simplest case, a subset of the skeleton's primary points can be used as three-dimensional browsing/auditioning cursors. This has the immediate benefit of providing polyphony, making it possible to reach toward multiple timbre regions at once, and pushes a basic exploration tool closer to becoming a musical instrument. A fundamental property of digital musical instruments is their ability to dynamically reassign pre-defined action-sound relationships (i.e., mappings), and here, nothing restricts the implementation of several different strategies that can be chosen freely during use.

The current system offers three navigation environments, which can be chosen by walking through one of three virtual "doorways" at a specific depth threshold within the physical tracking area. From the extreme rear of the tracking area and facing the sensor, walking forward to cross the depth threshold at the leftmost region imposes the simple multi-cursor mapping described above. The left and right hand are designated as active cursors, while distance between the hands and their individual three-dimensional delta values (i.e., accelerations) modulate parameters of various processing modules. Traversed in the other direction, the depth threshold is used to deactivate the mapping, freeing the user to cross it again at either the center or rightmost regions.

Mappings in the remaining doorways explore possibilities that arise when timbre spaces are grafted directly on the shoulders of the user. That is, rather than spreading audio grains throughout the entire tracking area, they are compressed to cubes attached to the users shoulders and auditioned based on the relative offset of the corresponding arm. Under this approach, a specific arm gesture activates roughly the same sequence of grains regardless of where the user stands in the tracking area. This means that the user's overall position can be used to select different chains of signal processing for application to the basic granular output. Leaning into specific regions, the user can choose to apply a network of flanging, pitch shifting, and pulsing at one moment, but ring modulation, filtering, and reverberation at the next. This embodied approach to timbre space navigation and audio processing provides access to a greater number of options, varies the orientation between user and space, and generally enhances large scale physical aspects of interacting with digital audio.

4.2. Open-air Fingertip Navigation

More nuanced control can be attained by browsing timbre spaces via open-air fingertip movements. Technically, this system relies on IR blob tracking, with reflective markers placed on tips of the thumb and middle finger of each hand. Because the markers are lightweight and passive (i.e., not powered), movement is not restricted. IR motion capture systems typically involve multiple cameras in order to capture data with three degrees of freedom. Here, the system is drastically reduced in comparison because the tracking area is relatively small, and portability, cost, and ease of use are top priorities. Nevertheless, it does provide very reliable tracking, including excellent depth resolution for three-dimensional tracking. Without additional cameras, spherical markers (which appear to be the same size from any angle at a given distance) are required in order to use IR blob size as an indicator of depth.

Rather than virtual doorways, pre-defined mappings are chosen based on which of the four fingertips enters a particular side of the tracking area first. A similar strategy was used effectively for an instrument described in [9]. As before, relative data between points can be used to modulate parameters of processing applied to the audio grains as they are browsed. For instance, by pinching with the left hand and rotating the wrist, the user can make specific adjustments to variables like delay time and pitch shift interval. The shape and size of the polygon defined by the four fingertips can be used for other layers of control. Considering the system as an instrument, we can say that its sound producing actions are extremely indirect, happening in relation to virtual objects that the performer must see to understand. With practice, strong relationships are formed between visual characteristics of the virtual elements and the resulting audio output.

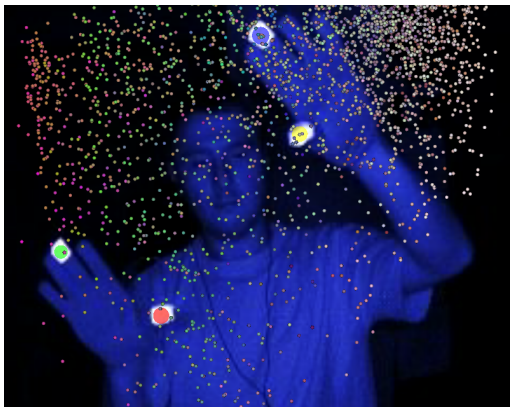


Figure 4: IR fingertip tracking for polyphonic timbre space browsing.

4.3. The Gesturally Extended Piano

The Gesturally Extended Piano (GEP) is an augmented instrument controller that exploits the pianist's arm movements for timbre space navigation and control over real-time transformation of the piano's acoustic sound. Among the most elementary pieces of movement information in the case of a pianist are the positions and angles of the forearms in relation to the keyboard. This information can be captured with IR blob tracking by following a minimum of two key points on each arm. As well as allowing different timbre spaces to be grafted onto the specified region of interest for polyphonic browsing with the four reflective markers as cursors, augmenting the piano with motion tracking enables intuitive control over sound characteristics that are usually inaccessible when playing the piano, such as continuous changes in pitch and volume.

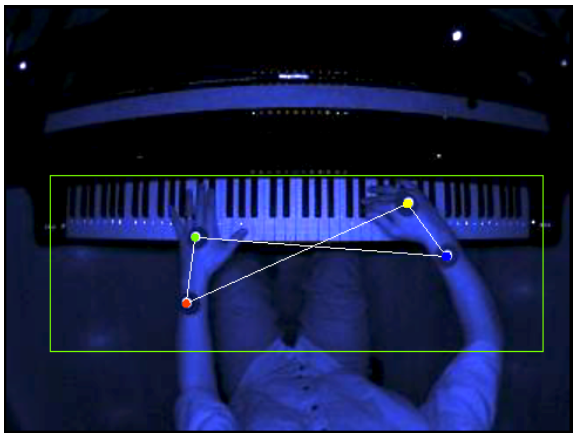


Figure 5: Overhead view of the GEP tracking system.

The GEP's camera and attached IR light array should be mounted directly over the keyboard in order to provide a clear overhead view of the entire playing surface and the pianist's arms. A convenient mounting point on grand pianos is the raised lid, but upright pianos can be fitted with the tracking system as well. The spherical reflective markers should be attached to the pianist's arms using a flexible silicone skin adhesive. Figure 5 shows the IR camera's view of the piano and user-defined region of interest, with red, green, blue, and yellow points drawn over top of the reflective markers, and connections drawn between some points. This animation provides useful feedback for the performer, and (as with the other systems described in this section) several interdependent control streams can be extracted from the scene.

Different mapping presets for the GEP controller can be selected based on entry conditions of the hands. For instance, the hands can enter from either the middle, far left, or far right of the region of interest, which provides three preset choices. The number of available choices can be doubled by observing whether the right or left hand is the first to enter each of these zones. Based on a depth threshold, the number of choices can be doubled once again, meaning that the pianist can choose to enter the region of interest either above or below the invisible threshold. This strategy avoids the need for any additional pedals or switches, keeping the amount of hardware to a minimum.

Space does not permit a detailed explanation of the mappings currently in use; however, one of the more intriguing options involves phase-vocoded scrubbing of a short audio buffer filled incrementally with a mix of desired audio fragments. This mapping relies on the distance between points on each hand, which can be lengthened or shortened by flexing the wrist forward or back. By defining a threshold, these motions can be used to trigger live audio capture into the buffer with the left hand, and clearing of the buffer with the right. The pianist can thus trigger the left hand before playing into the buffer, which is then scrubbed using the centroid of all four tracked points. Moving the hands between the low and high extremes of the keyboard, any particular moment of the sampled sound can be sustained by virtue of the phase vocoder, with further processing controlled via other aspects of arm orientation. After building up such a texture incrementally, the buffer clearing trigger of the right hand provides a means of bringing dense, sustained sound masses to a sudden and dramatic halt.

5. CONCLUSION

Both timbreID and DILib have been released under the GNU GPL as open source projects with the intention of further encouraging embodied approaches to digital exploration of sound relative to timbre. Though designed for native use in Pd, information generated by these libraries can be routed to any multimedia programming environment. Of the specific applications reviewed in Section 4, only the GEP has been used in live performance. After a period of experimentation, use, and refinement, software for these projects will be made available as open source tools for interested artists and performers.

6. REFERENCES

[1] G. Tzanetakis and P. Cook, "Marsyas: a framework for audio analysis," *Organised Sound*, vol. 4, no. 3, pp. 169–175, 1999.

[2] O. Lartillot and P. Toiviainen, "A matlab toolbox for musical feature extraction from audio," in *Proceedings of the 10th International Conference on Digital Audio Effects*, Bordeaux, France, 2007.

[3] J. Bullock, "Libxtract: A lightweight library for audio feature extraction," in *Proceedings of the International Computer Music Conference*, 2007.

[4] N. Collins, "SCMIR: A SuperCollider music information retrieval library," in *Proceedings of the 2011 International Computer Music Conference*, 2011, pp. 499–502.

[5] M. Puckette, "Pure data: Another integrated computer music environment," in *The 2nd InterCollege Computer Music Concerts*, 1996, pp. 37–41.

[6] Author, "A timbre analysis and classification toolkit for pure data," in *Proceedings of the International Computer Music Conference*, 2010, pp. 224–229.

[7] ——, "DILib: Control data parsing for digital musical instrument design," in *Proceedings of the 4th International Pure Data Convention*, 2011, pp. 176–180.

[8] M. Danks, "Real-time image and video processing in GEM," in *Proceedings of the 1997 International Computer Music Conference*, 1997, pp. 220–223.

[9] J. Oliver, "The MANO controller: A video based hand tracking system," in *Proceedings of the 2010 International Computer Music Conference*, 2010.

SPATIALIZED AUDIO FOR MIXED REALITY THEATER: THE EGYPTIAN ORACLE

Ajayan Nambiar

PublicVR,
333 Lamartine St., Jamaica Plain, MA 02130
ajayandn@gmail.com

Jeffrey Jacobson, Ph.D.

PublicVR,
333 Lamartine St., Jamaica Plain, MA 02130
jeff@publicvr.org

ABSTRACT

In the Egyptian Oracle, we project a simulation of an ancient temple onto a large projection screen. (See http://publicvr.org/egypt/oracle/shortvid.html.) We create the illusion of a contiguous space by matching the scale of virtual and physical objects. In the live performance, actors in front of the screen interact with human-operated avatar actors in the virtual space. As with any dramatic production, music, sound, and dialogue are a large part of the experience. Our goal is to create a unified aural space that extends from the physical through the virtual to encompass the entire performance. We use commodity electronics to produce an elegant affordable solution, which produces an impressive dramatic effect. We also confront fundamental issues typical of performances of this type, pointing the way to more advanced auditory solutions for interactive mixed reality spaces. This project was funded by the National Endowment for the Humanities, and the code is free to the public as open source.

1. INTRODUCTION

The Egyptian Oracle performance is a live reenactment of an authentic public ceremony from Ancient Egypt's Late Period. We project our Virtual Egyptian Temple on the wall at life scale, extending the physical theater into virtual space, as shown in Figures 1 and 2. The temple is a true three-dimensional space, which the audience navigates during scene changes. The central actor depicted in Figure 1 is a high priest (right), an avatar controlled by a live human puppeteer. The sacred boat (center) is another puppet, the oracle, which reveals the will of the temple god in the drama. Audience members represent the Egyptian populace acting out brief roles in the drama. By moving the boat, the Oracle has selected the woman on the left for a great honor. In other scenes, the priest interacts directly with audience members and a costumed live actress.

This experience is very difficult to understand from description alone. We highly recommend the video posted at http://publicvr.org/html/pro_oracle.html.

In the temple, ambient music sets the mood for each space, while moments of dramatic music and sound effects highlight the action, which the movie industry calls a "stinger." We create a sense of space with simple effects such as echo and reverb., which is adjusted depending on the current "location" in the space. For example, changes in reverb would immediately allow the audience to discern the transition from a big space to a small one or from an open space to a closed one. For greater aural continuity, the voice of the puppeteer, the live actress, and the currently selected audience member, are each channeled through a separate microphone to make all the voices part of the same auditory space. A live operator mixes the sounds, providing pleasing artistic balance and preventing problems such as feedback

We implemented the virtual environment and animations with the Unity game engine (http://unity3d.com) as an application that can run on a standard Windows® laptop. The software then introduces reverb with the aid of a 32-bit sound effects processor. It provides a wide range of effects such as echo, chorus, and double slap. The amplifier output can be increased from 80 watts through 2 channels to 130 watts through a 5-channel surround system. A powered amplifier along with a separate low-frequency line out gives us more bass control. A mixer gives a human sound-system operator more control over sound and eliminates floor noise.

Figure 1: Live people and avatars interact.

The resulting system provides a basic spatialization and is simple, portable, affordable, and effective. This work is a first step toward more advanced sound spatialization systems (e.g., http://www.vrsonic.com). The overall project was funded by the National Endowment for the Humanities, and Ajayan Nambiar produced the audio design for his Master's thesis

(Nambiar, 2011). The open source is available at the project website, http://publicvr.org/html/pro_oracle.html, and can be adapted to a wide variety of dramatic productions.

Several previous dramatic productions have used a sophisticated avatar/puppet for direct viewing by an audience. Ryu (2005) and her digital puppet performed a shamanistic drama for a live audience. Andreadis and his colleagues (2010) created a live performance by avatars/puppets in a virtual Pompeii, which was projected onto a large screen for a live audience. Anstey et al (2009) staged a number of dramas with a mixture of virtual and live actors. As with a traditional play, the audience is "along for the ride." The Oracle is unusual in its attention to spatialized audio, as we describe here.

2. VIRTUAL AUDIO SPACE

The virtual audio space is created in order to extend the imagination of the user beyond the boundaries of the physical space and create a controlled, realistic, and predictable aural environment (Figure 2). Our goal is to make the theater or classroom sound like an Egyptian temple, despite differences in venue. We are pursuing a basic, simple strategy first as we develop our approach.

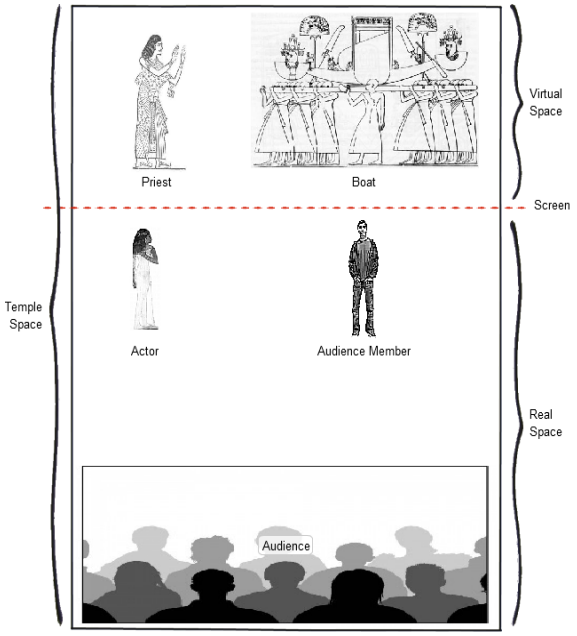


Figure 2: The unified visual and aural space, overhead view.[1]

The final audio consists of multiple layers starting from the ambient track produced within Unity3d. The next layer contains effects for realism and input from the audience. To preserve the continuity in space, the actor, puppeteer, and audience members are all provided with microphones to maintain a level field of sound. All the sounds are channeled through the mixer where the live sound-system operator can manually eliminate any interference. The mixer output goes to the amplifier to create the surround effect for the desired space.

Special sound effects help to further solidify the realism of the auditory space. Some of the effects are reverberations, echoes, Doppler effects, delays, panning, fade, and localizations. Sound effects can also help compensate for the shortcomings of the performance space. Manual control of sound effects is available in case the automatic effects fall short.

Stereo sounds lend themselves well to music but fall short when trying to simulate an environment because we perceive environments as three-dimensional, not just as left and right. We use surround sound (five speakers arrayed around the room), which helps the audience to perceive elements not actively displayed on the screen. Localization of sounds is much easier to emulate with surround sound, and our next step will be to add sound sources above and below the audience.

All of the electronic equipment for the show can be packed into the luggage shown in Figure 3. It includes a laptop computer, an Xbox 360 game controller and audio/mike headset for the puppeteer, a stereo amplifier, a mixer, two mobile microphones with their base station, 5 speakers with tripod stands, a short-throw projector, and cabling. The luggage also has room for physical props, the costume for the actress, and printed handouts. All of this is detailed in Nambiar (2011).

Figure 3: All the equipment needed for an Oracle show.

3. SONIC STRUCTURING

The goal of the Egyptian Oracle project is to provide realistic experience using a virtual space, and this paper describes the sonic/auditory dimension of that effort.

3.1. Sound Unification

The performance has multiple sources of sound input: (1) the voice of the puppeteer, (2) the voice of the live actress, (3) the voice of an audience volunteers playing small roles, (4) ambient music in the virtual temple, and (5) sound effects used to punctuate important moments in the action(stingers).

The sound inputs are fed to a high quality mixer, which requires a live operator, who balances the sounds for an artistically good effect. This is necessary because of the

[1] The image of the boat was taken from the Epigraphic Survey of Medinet Habu (1930-1940, Oriental Institute of Chicago) by the authors of http://ecuip.lib.uchicago.edu/diglib/social/ and copied into our diagram here.

unpredictability of the actors' voices and locations. For example, the volume of the speaker's voice changes, and the distance of the microphone from the actor and from the speaker varies. Also, the microphone input from the actor and puppeteer may have low-level audio hum present in the signal due to electrostatic or magnetic interference, which creates a noise floor. The noise is suppressed using the high impedance input of the mixer, producing a clean final output. Most importantly, if the actor's microphone is held at close proximity to the speaker, it results in an infinite input and output loop, which produces a high frequency sound through the speakers. The operator at the mixer can immediately drop the volume or mute the source to avert the issue. Malfunction of a device during a performance can lead to similar problems. There is no satisfactory way to automate the sound mixing, but we find the task straightforward and an opportunity for artistic judgment.

3.2. Sound Effects

The Egyptian Oracle software is built on the Unity3d game engine. Unity employs the FMOD sound engine (http://www.fmod.org/), capable of a variety of audio tasks. The Oracle incorporates triggers to handle these tasks with audio feedback. Unity's FMOD plugin can be used to regulate sound effects such as reverberations, echoes, delays, panning, fade-in, and fade-out. Since the ambient tracks are played from within the Oracle software, overlaying the desired effect helps set the conceptual space of the room in the mind of the audience. This also reduces work for the sound operator.

Finally, the sound operator can introduce variations manually by using the 32-bit effects on the mixer. The effects can apply to all sources collectively or to an individual sound channel at the discretion of the operator. For example, the actor's voice can be made to sound more emphatic, or the priest/puppeteer's voice deeper and more authoritative.

3.3. Spatialization

The Oracle software uses the game engine Unity, the sound library FMOD, and the sound system described to create a sense of space to help the audience feel as if they are inside the temple. To achieve true surround sound, the laptop being used must have an HDMI out or an optical out (SPDIF) port. The panoramic sound system allows us to localize sound sources within the soundscape. For example, the voice of the puppeteer appears to come from the side of the stage where the priest avatar is standing.

In this way, we employ spatialization to surround and enclose the audience, actor, and puppeteer within a single conceptual space, the Virtual Egyptian Temple. It blurs the line between the virtual and real worlds, including the audience within the performance.

4. MUSIC COMPOSITION

The pre-produced audio in the Oracle presentation consists of an ambient introduction, an ambient loop played in the background throughout the performance, and 14 tracks of "stingers" (short musical pieces to complement an action) and "traveling music" (music playing while the "camera" moves throughout the virtual space). Jon Hawkins wrote and produced the music and special effects in Logic Pro. See http://www.hawkinssounds.com and http://www.apple.com/logicpro/.

The ambient tracks consist of almost-static synthesizer drones and sounds, small chirps of birds, and slight wind (when the location is outside the temple). The deep synth drones provide a relaxing backtrack throughout the performance and reinforce the illusion of aural space in the virtual model. The "stingers" provide dramatic effect during actions at key moments in the drama. The "traveling music" provides a pleasant aural experience while the camera travels from one part of the temple to another during scene changes. The music tracks are designed to fit over the ambient backtracks or work as independent pieces as needed. They are also timed so that when they are triggered by an action in virtual space (i.e., the boat choosing an audience member), the stingers are synchronized with the animations.

Nobody knows what Egyptian music really sounded liked, because Egyptians had no musical notation. Many interpretations are possible, based on their surviving musical instruments and ethnographic evidence. Coptic Christian liturgy, for example, has elements that were bound to have come from Pharonic times. For this composition, however, we experiment with Hellenic elements and style because the Greeks, and later Rome, ruled Egypt for much of its Late Period, the setting for this drama.

When we secure the funding, we will record live performances of reconstructed ancient instruments. For now, we are using electronic simulations consisting of samples, filter effects, EQ, and harmonic manipulators to imitate the sounds of popular ancient instruments: the kithara (an ancient harp/lyre), pan flute (a wooden multi-chambered flute), and a variety of percussive instruments (drums, bells, finger cymbals, and shakers).

We developed the sounds by feel, using our artistic judgment. We tested and refined it in a variety the venues of different sizes and acoustic properties, and find that it works well. The next step will be to use more advanced software to simulate the acoustic properties of the virtual temple as they might have been in real life.

5. APPLICATIONS

The Egyptian Oracle performance has a dual purpose – to demonstrate the potential of mixed reality theater and to educate the public on a key aspect of ancient Egyptian culture that the public is not likely to have seen elsewhere. Religious performance and ritual permeated ancient Egyptian culture, and it is related to much of the ceremony in the Abrahamic religions (Judaism, Christianity, and Islam). The current version of the Egyptian Oracle performance was originally designed for children 10 to13 years old and for family audiences, but it has been well received by adult audiences as well. The performance is currently well suited to special showings at community theaters, K-12 schools, and science museums. As we develop it further, we will add depth to the narrative and refine the artwork. Our goal is to distribute the Egyptian Oracle to museums in the humanities in the form of a documentary film and online as a distributed virtual world. The spatialized audio described in this paper is a first step to harnessing higher

fidelity audio displays, primarily for audiences in museums and large dome (planetarium) venues. Obviously, the technology and approach could be used for educational theater on a wide range of topics.

6. CONCLUSION

The Egyptian Oracle sound spatialization project began with a purely software-based solution in mind. But performances at different venues revealed that a purely software-based solution was incapable of handling all the demands of live sound. In response, we devised a hardware solution based on a live sound system operator working with the mixer. While a live operator working alone can exercise more judgment and far greater flexibility than any automated system, the cost is greater than that of a turnkey solution. In the end, our solution provides a real and elevated sound experience for an excellent visual depiction of the Egyptian Oracle.

7. REFERENCES

Andreadis, A.; Hemery, A.; Antonakakis, A.; Gourdoglou, G.; Mauridis, P.; Christopolis, C.; and Karigiannis, J. N. (2010). *Real-Time Motion Capture Technology on a Live Theatrical Performance with Computer Generated Scenery,* 14th Panhellenic Conference on Informatics. IEEE Computer Society, ISBN: 978-0-7695-4172-3 http://doi.ieeecomputersociety.org/10.1109/PCI.2010.14

Anstey, J.; Patrice Seyed, A.; Bay-Cheng, S.; Pape, D.; Shapiro, S. C.; Bona, J.; and Hibit, J. (2009). *The Agent Takes the Stage*, International Journal of Arts and Technology 2009 - Vol. 2, No.4, 277-296.

Nambiar, A. (2011). Sound Spatialization For the Egyptian Oracle, Master's thesis for a degree in Professional Studies, Department of Digital Media, Northeastern University, MA, USA. http://publicvr.org/publications/NambiarA2011.pdf

Jacobson, A. (2011). *Egyptian Ceremony in the Virtual Temple: Avatars for Virtual Heritage*, Whitepaper and Final Performance Report to the National Endowment for the Humanities. Digital Startup Grant #HD5120910, 2010-2011 academic year. http://publicvr.org/egypt/oracle/whitepaper.pdf

Ryu, S. (2005) *Virtual Puppetry and The Process of Ritual,* Computers and Composition (C&C.): Elsevier, 2005.

8. FINAL NOTE ON ICAD 2012

This short paper will be presented during the poster session at the International Conference on Auditory Display in Atlanta, June, 2012. We will demonstrate the software and a stereo version of the sound (using headphones) for passersby. In the future, we would like to stage the full performance, but that is not possible at this time.

ACOUSTIC INTERFACE FOR TREMOR ANALYSIS

David Pirrò[1], *Alexander Wankhammer*[1], *Petra Schwingenschuh*[2], *Alois Sontacchi*[1], *Robert Höldrich*[1]

1 - Institute of Electronic Music and Acoustics,
University of Music and Performing Arts Graz, Austria
2 - Division of Special Neurology, Medical University of Graz, Austria
pirro@iem.at, wankhammer@iem.at, petra.schwingenschuh@medunigraz.at
sontacchi@iem.at, hoeldrich@iem.at

ABSTRACT

In this paper we introduce new methods for real-time acoustical tremor diagnosis. We outline the problems of tremor diagnosis in the clinical context and discuss how sonification can complement and expand the existing tools neurologists have at their disposal. Based on three preliminary sonification experiments upon recorded tremor movement data, we show how temporal as well as spectral characteristics of tremor can be made audible in real-time.

Our first observations indicate that differences among tremor types can be made recognizable via sonification. Therefore, we suggest that the proposed methods could allow for the formulation of more confident diagnoses. At the end of the paper, we will also shortly outline the central topics of future research.

1. INTRODUCTION

Tremor is the most common movement disorder. It is defined as a rhythmic and involuntary oscillation of a body part, caused by reciprocal nervous innervations of muscles [1]. The wide spectrum of tremor forms is summarized in the 1998 consensus statement of the Movement Disorder Society [2], whereas the most common forms include the essential tremor, parkinsonian tremor, dystonic tremor and psychogenic tremor. It is well known that each different tremor form can be the symptom of a specific disease [1]. Therefore, reliable classification and quantification of different tremor types is of strong clinical interest. A correct tremor diagnosis early in a diseases course is crucial in order to provide adequate treatment and medication for the patient.

In many cases a confident clinical diagnosis mainly based on the visual analysis of a tremor by neurologists experienced in movement disorders is possible. Nevertheless, these neurologists have to be highly specialized in this form of diagnosis and in some situations uncertainty remains. Therefore, further investigations based on structural and functional imaging, video analysis, accelerometry and other electrophysiological investigations can be necessary. Although such methods offer important additional information for a final diagnosis, the ex-post analysis and interpretation of recorded data is typically very time-consuming and hard to implement in the daily routine of clinical examinations. Besides, these methods do not support the neurologist during the personal contact with the patient.

The sonification experiments presented in this paper aim at extending established tremor analysis methods by an acoustical interface for tremor diagnosis. Based on real time sonification of acceleration data, detailed information on the temporal as well as spectral characteristics of tremor could be made audible to the neurologist while interacting with the patient. As sonification could provide an additional modality to perception, it would allow for a holistic analysis of the observed tremor avoiding the major drawbacks of ex-post analysis methods.

2. AUDITORY DISPLAY FOR MOVEMENT DATA

Sonification of movement data in the medical context is being employed in different areas and in conjunction with various motion capturing technologies. For example in virtual rehabilitation [3], sonification provides objective real-time information for analysis. In physiotherapy, sonification has been used to offer clear feedback for therapists and patients during rehabilitation exercises, e.g. with the sonification of EMG (Electromyogrphy) data [4]. Further, the auditory channel can be employed to augment the perception and proprioception of the subjects to heighten their motivation [5]. From a more general perspective, auditory data displays also gain increasing interest in multimodal biomedical data representation [6], caused by the growing number of simultaneous data streams that have to be perceived and analyzed. However, the sonification of tremor movements as a diagnostic tool is a novel research topic that has not been addressed until now.

The sonification studies we present here concentrate on tremor in Parkinson's disease, essential tremor, and psychogenic tremor. From a medical point of view these are clinically sometimes difficult to distinguish and therefore a clear discrimination by means of acoustical tremor analysis would be of great importance.

Since sonification will serve as a tool for neurologists, not sound specialists, our principal aim is to make differences between tremor types perceivable as clear as possible. To avoid auditory information overload, we try to lower the complexity of the sonification, by associating only the most significant and well-defined qualities of the data with distinct sound attributes. At the same time, we try not to oversimplify but to preserve all relevant information present in the data.

The preliminary experiments presented in this paper have been carried out on pre-labeled tremor data that has been captured by one of the authors during previous clinical studies. To evaluate the quality of the proposed sonification methods as diagnostic tools, a prospective clinical study with multiple neurologists who have been trained with the sonification system will be carried out at the Medical University of Graz in 2012. The neurologists will be asked to classify 30 different tremor patients with known diseases (approx. 10 per tremor form), basing their diagnoses solely on audio files representing sonifications of recorded

tremor data. Although such an "audio only" restriction does not reflect a real-world situation, it allows to asses the quality of the proposed method, when compared to the established standard of clinical tremor diagnosis.

3. SONIFICATIONS

In the following sections, we describe the three sonification experiments carried out on the data set we have at our disposal. At the end of each section, we briefly outline how the different tremor forms could be distinguished by the respective sonification approach.[1]

The acceleration data we work with has been captured with a sampling rate of $f_s = 1\ kHz$ using a 3-axes accelerometer[2] taped to the backside of the proximal phalanx of the index finger of a patient's hand. As the typical frequency range of pathological tremor lies approximately between 3 and 15 Hz, we apply a DC removal filter and a low pass filter with a cutoff frequency of $f_{cL} = 70\ Hz$ to the acceleration signals and upsample the data to $44.1\ kHz$. When the system will later be integrated as real-time diagnostic tool, the data will be captured using the same sensor, but already at audio rate. Therefore, real-time capability of the designed system has already been a crucial requirement for our preliminary studies.

3.1. Frequency-Shifted Audification

In our first approach we aim at translating the acceleration data into sound in a simple and direct way i.e. without any sophisticated pre-processing. As the signals we are confronted with exhibit frequencies mostly below the audible range and our system has to be real-time capable, direct data audification (e.g. transposition via sample rate conversion) is not a viable option. However, as the ear is very sensitive to changes in amplitude and frequency, ranging from loudness fluctuations to different forms of roughness, the tremor signals turned out to be especially well-suited to serve as modulators of fixed frequency carriers. To highlight the rhythmic structure of the observed tremor signals, we apply simultaneous amplitude (AM) and frequency modulation (FM) to the carrier signals. Since maxima in frequency coincide with maxima in amplitude of the resulting modulated signal, rhythmic or dynamic changes of the tremor become clearly audible.

The following formula describes this sonification approach for one axis:

$$x_{son}(n) = \hat{x}(n) \sin\left(2\pi n(f_x + kx(n))\right) \quad (1)$$

where $\hat{x}(n)$ represents the half-wave rectified acceleration signal $x(n)$ along the x-axis to avoid a doubling of the perceived modulation frequency in relation to the observed tremor movements and the frequency f_x of the carrier is fixed. The amount of FM can be controlled by the modulation index k, resulting in a pure amplitude modulation with suppressed carrier when $k = 0$. The signals $y_{son}(n)$ and $z_{son}(n)$ can be computed in the same way, but with different carrier frequencies f_y and f_z. The simultaneous sonfication of all three axes can then be defined as follows:

$$tot_{son}(n) = a_x x_{son}(n) + a_y y_{son}(n) + a_z z_{son}(n) \quad (2)$$

where the amplitudes a_x, a_y and a_z can be controlled separately, allowing to isolate single acceleration axes or planes for the sonification; the squared sum of these weighting parameters is normalized to one.

Though quite basic, this approach allows us to rapidly explore the data and get a glimpse of how the different tremor typologies can be characterized. In particular, a first distinction between tremor types can be based on their temporal characteristics. While the parkinsonian tremor shows a regular pulsation that can remain steady for long time intervals ($10 \sim 20$ seconds), there seems to be no regularity of any kind in most essential tremor cases as the sine is modulated by a quite noisy signal. In psychogenic tremor, pulses appear for short time intervals, but the beats do not present a steady repetition rate; they seem to falter, generating hesitating rhythms.

3.2. Spectral Features

As amplitude and frequency of a tremor are two very important parameters for the detection and quantification of tremor types, spectral analysis of recorded accelerometry and EMG data has been used widely by neurologists [7, 8]. Typically a spectral representation of the investigated tremor signal (e.g. the power spectrum) is computed based on the Fourier transform, followed by the extraction of specific spectral descriptors. Since many neurologists are familiar with these descriptors we want to make them audible in a real-time sonification.

The most commonly used parameters in the context of human tremor analysis are the peak tremor frequency, the total power of the spectrum between 1 and 30 Hz and the half-width power, where the half-width is defined as the frequency interval between the two values left and right to the main peak, at which the spectral power density is half of the peaks' power (see Figure 1). It is important to note that for clinical ex-post analysis, these features are typically computed over relatively long observation periods (e.g. 30 seconds), which practically eliminates any temporal information inside the signal.

When analyzing the power spectra of different tremor forms, we can basically distinguish three different scenarios (see Figure 1): a nearly harmonic spectrum, mostly in the Parkinson's disease (top); no narrow or clear peaks, but a broader region in the lower part of the spectrum, recurrent in the essential tremor (middle); only one prominent peak, frequent in the psychogenic tremor (bottom). To parametrize the sonification algorithm, we therefore decided to use features similar to those depicted in Figure 1. We extract the central frequency of the main peak (if present), its half-width power and we detect the presence of side peaks or harmonics.

As the relevant tremor frequencies lie in a very low and narrow part of the spectrum, we have to perform the real-time spectral analysis inside sliding windows of at least one second, in order to achieve the necessary frequency resolution. That way, the temporal structure of the signal is preserved, but its level of detail is limited to the selected window length.

In the sonification, where each axis can be sonified individually, we use a Karplus-Strong [9] algorithm. The excitation signal of the algorithm is pink noise and the base frequency is proportional to the central frequency of the main peak. The location of the main peak is determined by parabolic interpolation between neighbouring bins: this way, "jumps" of the base frequency are avoided when the the main peak of two consecutive frames resides

[1] Examples: http://iem.kug.ac.at/index.php?id=13661

[2] Sensor details: http://www.biometricsltd.com/accelerometer.htm

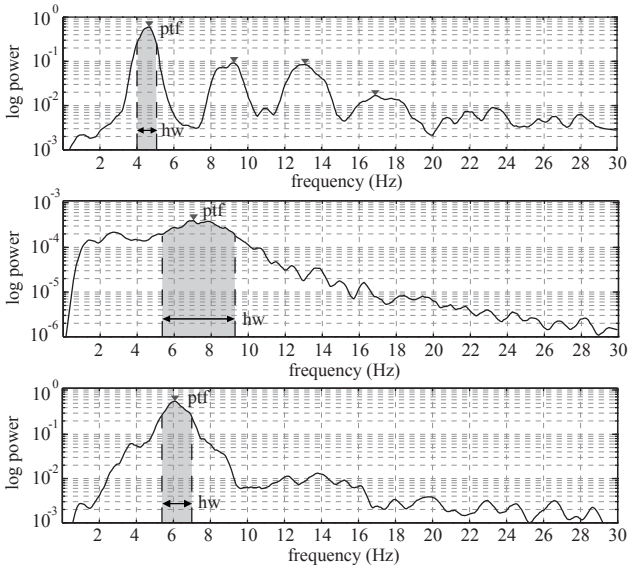


Figure 1: Typical power spectra of tremor signals with indicated peak tremor frequencies (ptf), half widths (hw) and side peaks.

in different bins. The width of the main peak is used to control the feedback factor of the delay line in the algorithm. If the peak is narrow, the feedback factor is nearly equal to 1.0. Instead, if the peak gets broader, the feedback factor is diminished and gets down to 0.0 when no clear peaks are detected. The output is then passed through a low pass filter: if harmonics of the main peak are detected, the cutoff frequency of the filter is adjusted to be eight times the base frequency and set to be equal to it if no significant side peaks appear in the spectrum.

The resulting sound presents a clearly pitched tone, if the spectrum exhibits one or multiple peaks. When sonifying acceleration data of a parkinsonian tremor, the spectrum of the tone has more overtones and the pitch remains quite stable over longer time intervals. On the contrary, when sonifying psychogenic tremor data, the generated tone can have more noise mixed in. It has a tighter spectrum and its base frequency moves quite frequently. Essential tremor generates a more noisy sound from which occasionally tones pop out but immediately disappear.

3.3. Translation and Rotation

For the final sonification approach we analyze the spatial movement pattern of the trembling hand. Since a sonification based on detailed information on the exact hand movement would presumably provide too much auditory information, we developed a method to separate the observed motion into its major components.

Assuming that the main components of a hand movement are typically located on a slowly changing plane in space, we project the three dimensional acceleration vectors onto this plane, in the following referred to as the "plane of movement". This projection does not only reduce the dimensionality and amount of data we have to process, but also offers important information on the investigated motion.

To identify the plane of movement of a motion in real-time, we have to detect its major acceleration components during short observation periods. As the spatial spread of successive acceleration vectors directly represents the amount of acceleration into the respective directions, Principal Component Analysis (PCA) is a suitable method to identify the two main axes of acceleration. PCA basically detects the direction of the greatest variance of the data and places a first axis in this direction (the first principal component). The next axis (second principal component) is chosen perpendicular to the first axis along the direction of the next greatest variance. Hence, the first two principal components directly represent the vectors defining the plane of movement.

As we have to process the captured data in real-time, the principal components are computed based on an iteratively updated covariance matrix $\mathbf{\Sigma}(n)$. The first two eigenvectors $[\boldsymbol{\gamma}_1(n), \boldsymbol{\gamma}_2(n)]$ of $\mathbf{\Sigma}(n)$ represent the principal components that are used to project each three dimensional input sample $\boldsymbol{a}(n) = [x(n), y(n), z(n)]^T$ onto the plane of movement. The resulting transformed input samples $\tilde{\boldsymbol{a}}(n)$ are called score vectors.

$$\tilde{\boldsymbol{a}}(n) = \begin{bmatrix} \tilde{a}_1(n) \\ \tilde{a}_2(n) \end{bmatrix} = [\boldsymbol{\gamma}_1(n), \boldsymbol{\gamma}_2(n)]^T \boldsymbol{a}(n) \tag{3}$$

Successive score vectors now define a two dimensional trajectory that offers important information on the amount of translation and rotation inherent to the observed motion. This becomes clear, when we analyze the characteristic data distributions related to purely translational and rotational movements: a translational movement will lead to a "line-like" distribution of acceleration values, as only the sign and magnitude of the acceleration vectors changes over time, while a rotational movement will lead to a "circle-like" data distribution, as only the direction of the acceleration vectors constantly changes.

Considering these characteristic distributions projected onto the plane of movement, we can now make two important observations: any progression along the axis defined by the first principal component can be caused by rotation and translation, while changes along the second PCA axis can only be caused by rotational components. Therefore, the second element of the score vector directly represent the rotational signal component $r(n)$ of an observed motion. To get a definition of the translational component $t(n)$, the influence of the rotational component has to be removed from the first element of the score vector. Hence, after calculating the Root Mean Square (RMS) of each element of the score vector, the translational component can be obtained as follows

$$t(n) = \frac{RMS\{\tilde{a}_1(n)\} - RMS\{\tilde{a}_2(n)\}}{RMS\{\tilde{a}_1(n)\}}\, \tilde{a}_1(n) \tag{4}$$

In the sonification we use the translational and rotational components, $t(n)$ and $r(n)$, and the smoothed sum $s(n)$ of the x, y and z components of the acceleration signal.

$$t(n)HPF\{s(n), 1000\} + r(n)s(n)sin(2\pi fn) \tag{5}$$

In the first part of the sonification we pass the signal $s(n)$ through a second order high-pass filter $HPF\{s(n), 1000\}$ with a cutoff frequency of 1000 Hz and multiply the result with the translational component; this generates clicks or sort of thumping beats. In the second part we use it to modulate the amplitude of a fixed frequency sine and multiply the result with the rotational component.

This way, we try to sonically separate and enhance translational or rotational qualities of movements by associating them

with two contrasting sound qualities that can be easily distinguished. Further, as the signal $s(n)$ contains all the temporal details of the movement itself, we do not loose this information in the sonification and can rely on it when making distinctions between the three tremor forms.

Applying this analysis and the related sonification to the acceleration signal, we can point out some important observations. In the essential tremor, the rotational component is most pronounced; even if the movement is highly disordered and not regular, the major component is typically rotational, as the irregular clicks generated by the translational component remain in the background. The psychogenic tremor is mostly characterized by a strong translational component; the rotational component can also be present and even dominate, but only for short time intervals. In parkinsonian tremor, both components can be present, but in most cases clear and regular beats can be heard, helping to identify this tremor.

4. DISCUSSION

Our first experiments showed that sonification could indeed be a promising extension to the diagnostic tools already used by neurologists. In particular, the temporal qualities of tremor movements, which are transported by all our sonification approaches, seem to bear important information that can be crucial in the distinction between the various forms. Besides, our approach to separate translational and rotational movement components is a novel analysis method and could also be an interesting step forward in clinical tremor research.

Even if most of the tremor forms can be identified via the sonification, some cases remain unclear and a certain amount of ambiguity is inherent to all of the sonifications we presented. Although complementing one sonification with another can sometimes be useful, it often leads to more confusing results, as it packs too much information into one sound.

Considering the quite limited set of data examined so far, the new recordings that will be made in the next months will help us to sharpen the tools we created. Still, a different analysis approach that could eventually give us a more holistic view of tremor would be desirable. In the next section we will therefore introduce an analysis method that could possibly meet our needs in this respect.

5. OUTLOOK

As the preliminary evaluation results are very promising, we are planning to extend the current system with a more sophisticated data analysis method. In particular, the correct separation of simultaneous movement patterns or tremor modes inside overlapping frequency bands would offer valuable information for the sonification process.

Depending on the investigated tremor type, the frequencies of individual tremor modes may significantly vary throughout a tremor recording. When using traditional spectral analysis methods (see section 3.2), it is often impossible to determine if different peaks inside a spectrum represent the coexistence of separate tremor modes, one mode residing in multiple frequency bands or if they are caused by local oscillations during the observation period.

To overcome these difficulties, recent studies [10] propose to use empirical mode decomposition (EMD) [11], a relatively new time-frequency analysis method for nonlinear and non-stationary data, for the analysis of tremor signals. Unlike Fourier analysis, where signals are assumed to be a composition of linear, stationary components, EMD decomposes any arbitrary time series into a set of superimposed oscillations (AM/FM modulated signals), called intrinsic mode functions (IMF). Practical investigations on tremor data have shown that individual IMFs carry important information on the investigated tremor movement, as they adaptively follow the nonlinearities and non-stationarities inside the signal.

Considering our observations presented in section 4, we are planning to apply EMD not only to the three dimensional accelerations vectors, but also to the rotational and translational signal components. The resulting IMFs could then serve as new input parameters for a temporally as well as spectrally detailed sonification approach.

6. REFERENCES

[1] K. Wyne, "A comprehensive review of tremor." *JAAPA*, vol. 18, no. 2, pp. 43–50, 2005.

[2] G. Deuschl, P. Bain, and M. Brin, "Consensus statement of the movement disorder society on tremor," *Movement Disorders*, vol. 13, no. S3, pp. 2–23, 1998.

[3] G. Burdea, "Keynote address: Virtual rehabilitation- benefits and challenges," *1st International Workshop on Virtual Reality Rehabilitation VRMHR*, 2002.

[4] S. Pauletto and A. Hunt, "The sonification of emg data," *Proc. of the 12th ICAD*, 2006.

[5] K. Vogt, D. Pirrò, I. Kobenz, R. Höldrich, and G. Eckel, "Physiosonic - movement sonification as auditory feedback," *Proceedings of the 15th International Conference on Auditory Display*, 2009.

[6] E. Jovanov, D. Starcevic, V. Radivojevic, A. Samardzic, and V. Simeunovic, "Perceptualization of biomedical data. an experimental environment for visualization and sonification of brain electrical activity," *Engineering in Medicine and Biology Magazine, IEEE*, vol. 18, no. 1, pp. 50 –55, jan.-feb. 1999.

[7] J. Timmer, M. Lauk, and G. Deuschl, "Quantitative analysis of tremor time series," *Electroencephalography and Clinical Neurophysiology/Electromyography and Motor Control*, vol. 101, no. 5, pp. 461–468, 1996.

[8] P. Mansur, L. Cury, A. Andrade, A. Pereira, G. Miotto, A. Soares, and E. Naves, "A review on techniques for tremor recording and quantification," *Critical Reviews™ in Biomedical Engineering*, vol. 35, no. 5, pp. 343–362, 2007.

[9] K. Karplus and A. Strong, "Digital synthesis of plucked string and drum timbres," *Computer Music Journal*, vol. 7, no. 2, pp. 43–55, 1983.

[10] E. Rocon de Lima, A. Andrade, J. Pons, P. Kyberd, and S. Nasuto, "Emd: A novel technique for the study of tremor time series," in *World Congress on Medical Physics and Biomedical Engineering 2006*. Springer, 2007, pp. 992–996.

[11] N. Huang, Z. Shen, S. Long, M. Wu, H. Shih, Q. Zheng, N. Yen, C. Tung, and H. Liu, "The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis," *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, vol. 454, no. 1971, p. 903, 1998.

CAPTURING AUDIENCE EXPERIENCE VIA MOBILE BIOMETRICS

Yuan-Yi Fan

University of California Santa Barbara
Media Arts and Technology
3309 Phelps Hall Santa Barbara California 93106
dannyfan@mat.ucsb.edu

René Weber

University of California Santa Barbara
Department of Communication
4405 SS&MS Building Santa Barbara California 93106
renew@comm.ucsb.edu

ABSTRACT

Different from computer vision based approaches in audience participation research, such as in Glimmer [1] and in Flock [2], this paper presents a mobile approach to collecting and visualizing bodily responses from audience members. A mobile biometric application is designed as a novel medium that interfaces audience members to experienced content. To realize our goal on a mobile platform, a combination of video-imaging-based heart rate measurement and Zeroconf networking technology [3] (Bonjour) is implemented. As a proof of concept, we successfully collect continuous heart rate values from 3 mobile phones devices simultaneously and use the derived heart rate statistics to drive artistic audio and visual rendering. Preliminary results include two iOS applications and two mobile-biometric-enabled media arts installations.

1. INTRODUCTION

Similar to the use of biometrics in electronic art [4], a novel mobile biometrics system is designed and implemented in this paper. In designing public interactive interfaces in settings like theatres, galleries, theme park, and museums [5], our mobile biometrics provide a new design parameter that captures audience or spectators' bodily response. We first review the use of technology in audience participation and response research in the fields of affective computing [6], electronic arts [7], and collaborative musical experiences [1][2][8]. Then, we propose the use of a mobile biometric application as a probe to measure audiences bodily responses and demonstrate electronic artistic applications we implemented based on our mobile biometrics. Third, the design and implementation is described. Fourth, we present a preliminary evaluation of our mobile heart rate measurement implementation using a commercial Photoplethysmograph sensor (Biopac System Inc. [9]). Finally, future research directions and final thoughts are discussed.

2. CAPTURING AUDIENCE EXPERIENCE

Tools that allow participation of large audiences in electronic art applications have become an emergent field of research [8]. A real-time response device for collecting listeners' impressions on a temporal arts piece has been discussed in [10]. The type of biometrics we choose to implement in this study is a mobile heart rate monitor using the built-in camera of mobile phone devices. Our implementation is non-invasive because we are using an optical signal. Mobile biometrics and imaging-based biometrics, particularly heart rate measurement, have recently become emergent in both academic and in commercial circles due to recent advancement in software design and the miniaturization of the necessary hardware [6] [11]. The issue of scalability has been addressed in designing such a system for interactive audience participation [12]. We intend to introduce our mobile heart rate monitor as valid indicator of audience members arousal states. The main advantage of our mobile biometrics is its simple deployment and use in public interactive multimedia installations. This paper aims to explore the design and validity of such an unobtrusive, real-time audience participation system

2.1. Design

Existing biometrics-based electronic art installation have been restricted to either wired physiological data acquisition equipment [6] or limited to only one individual, as opposed to audience members in a real-time and aggregated fashion [25][26]. Application in the current affective computing field has personal heart rate monitoring devices on mobile platforms [10][11], which inspires us to extend such optical sensing approaches with network streaming capability. Our overall design objective is to create a tool that enables experiential design in the context of electronic art that takes into account the bodily responses of the audience members. The scenario we consider is a setting such as an interactive installation or an electronic art concert that uses audience participation or audience response technique as a design parameter. Specifically, the use of available mobile phone devices will provide a sense of feedback and control for audience members which is likely to increase the audiences participation in the event. To achieve such a design objective, a mobile biometric application that is easily deployable and measures the audience members' heart rate unobtrusively and in real-time is required.

2.2. Implementation

The system implementation consists of three parts. First, an iOS application utilizes the built-in camera of the mobile device in measuring the users heart rate. Second, Bonjour is implemented so that the network connection can be set up with minimal configuration steps. After network connection is set up, the heart rate value can be transmitted back to server via OpenSoundControl (OSC) [13] in real-time. Last, a server-side application capable of receiving heart rate values from audiences mobile device is implemented in a Max/MSP [14] environment. For the heart rate measurement, audience members have to place the index finger of their dominant hand on the mobile devices camera lens (see Figure 2). Thereafter, the iOS application accesses each video frame

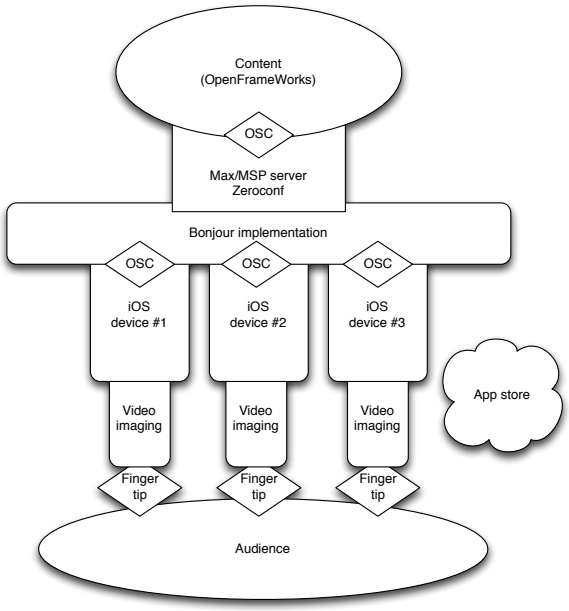


Figure 1: System design

from the built-in camera, and uses a finger-blood-volume-pulse signal in combination with a heart rate detection algorithm similar to that implemented in the Heartphone project [9] to compute the audience members heart rate. To enhance the performance of the heart rate measurement on a mobile device, we use the devices flashlight when acquiring the finger-blood-volume-pulse signal is in acquisition. To realize Zeroconf networking implementation on an iOS platform, Bonjour is implemented in our application. This makes it possible to connect to a server in a local area network with one-click on the user interface and it greatly reduces the configuration steps in setting up the server IP and port number. Finally, it uses our mobile biometrics to scale a 3D animation rendering as an example of the technologys capability. Intuitively, a 3D human heart model is animated based on the computed heart rate statistics. Visualization is done using OpenFrameWorks C++ Toolkits [15]. Our prototype system is implemented using a laptop that runs Max/MSP as server and three iOS mobile phone devices (iPhones) as clients. As our first prototype, the system is currently restricted to work within a Local Area Network (LAN), but later implementations will make it possible to deploy the system via an internet protocol.

3. PRELIMINARY RESULTS

As a working prototype of our system, the Max/MSP application running on the server successfully collects continuous heart rate values from 3 iPhones and uses the computed statistics for artistic graphics rendering. In Figure 3, we show an application that scales a human heart model based on incoming heart rate values from our mobile biometric application. So far, our mobile biometrics have been used in two iOS applications and two public media arts installations. The two iOS artistic applications are BioCymatics and HRclient, and the two public media arts installation are Ambient

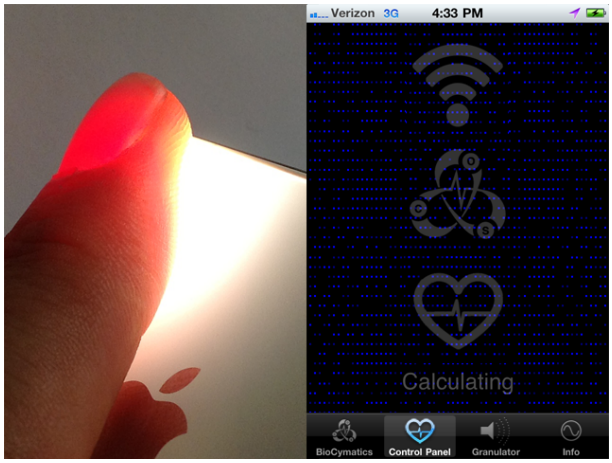


Figure 2: The image on the left illustrates where user should place his/her index finger for the heart rate measurement. The image on the right shows the user interface of our mobile biometric application.

Vision and Fight Or Flight.

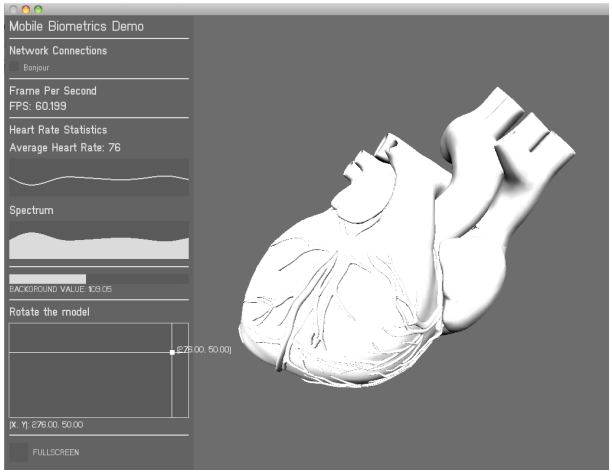


Figure 3: Intuitive 3D heart model is animated based on input signal from our mobile biometrics application. The control interface is implemented using OpenFrameWorks addon ofxUI [16]

The BioCymatics app explores the artistic use of biometric feedback signals (see Figure 2), such as using heart rate values to drive the graphical rendering of Cymatic patterns [17] and granular sound synthesis [18, 19]. In Figure 4, Ambient Vision [20] is an interactive audiovisual installation that addresses rippled mental images as the product of perceived stimulus and the internal bodily responses. The internal bodily responses refer to the heart rate collected using our mobile biometric iOS app HRclient while the external perceived stimulus is reconstructed based on information from Microsofts Kinect sensor. Throughout the installation, all software is configured remotely. During the exhibition, the spectator could easily participate in the exhibition by download-

ing the HRclient from app store freely. The above two examples demonstrate the design of our mobile biometrics application for both application designer and spectator participation in electronic art application. Fight Or Flight [21, 22] is another public installation that uses HRclient app. When the collected heart rate value from HRclient app exceeds certain threshold, it triggers a Boid swarming algorithm [23] to change between the calm state and chaotic state.

Figure 4: Interactive installation that uses the 3rd version of the mobile biometric application. Ambient Vision at Collider media arts series exhibition, Akron, Ohio, March 29-31, 2012 [20]

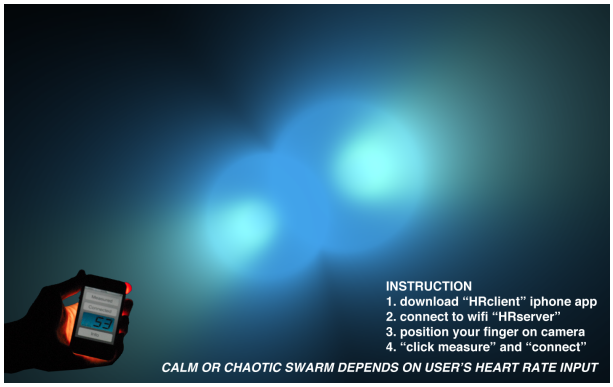


Figure 5: Interactive installation that uses the 3rd version of the mobile biometric application. Fight or Flight at UCSBs PRIMAVERA of Contemporary Arts and Digital Media, April 9-12, 2012 [21]

4. DISCUSSION AND FINAL THOUGHTS

Feedback from public installations that use our mobile biometric application is generally positive. In setting up Ambient Vision at Collider exhibition, we received good feedback from gallery staff for that we didn't have to ship bio-sensing equipment to the installation site. Since we deployed our mobile biometric application via the App Store, it reduced the potential complex software and hardware configuration as well as the equipment shipping insurance. One common negative feedback from user was our heart rate

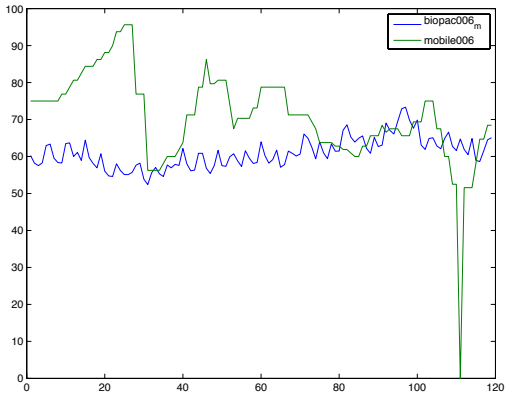


Figure 6: The blue line represents heart rate data collected using a commercial Photoplethysmograph sensor (Biopac System Inc. [9]), and the green line represents data collected using our mobile biometric application.

measurement is sensitive to motion artifact and background ambient light conditions. Although heart rate signal has long been used in the electronic art [7] and sonification field [24], our mobile biometric application is novel for its networking capability and ease of use in measuring heart rate. The system described in this paper enables research in techniques for aggregating audience input [12] and in other facets of the audience experience [5], such as interactive spectator and performer awareness. Future works involve the improvement of the heart rate measurements accuracy, large-scale installation based on our mobile biometric application, exploring the use of bodily responses in designing audience experience, and the use of mobile continuous self-reports in combination with our mobile heart rate monitor.

5. ACKNOWLEDGMENT

The author thanks Charlie Roberts for help with the iOS development.

6. REFERENCES

[1] J. Freeman, *Glimmer for chamber orchestra and audience*, PhD dissertation, Columbia University, 2005.

[2] J. Freeman, "Creative collaboration between audiences and musicians in Flock," *Digital Creativity*, Vol. 21, No.2 pp. 85-99, 2010.

[3] `http://en.wikipedia.org/wiki/Zero_configuration_networking`.

[4] R. McGee, Y.Y. Fan, and S.R. Ali, "BioRhythm: a Biologically-inspired Audio-Visual Installation," in *Proceedings of the International Conference on New Interfaces for Musical Expression*, University of Oslo and Norwegian Academy of Music, 2011, pp. 80-83.

[5] S. Reeves, S. Benford, C. OMalley, and M. Fraser, "Designing the Spectator Experience," in *Proceedings of the SIGCHI*

conference on Human factors in computing systems CHI 05, ACM Press, New York, 2010, pp. 741.

[6] M.Z. Poh, K. Kim, A.D. Goessling, N.C. Swenson, C. Nicholas, and R.W. Picard, "Heartphones: Sensor Earphones and Mobile Application for Non-obtrusive Health Monitoring," *2009 International Symposium on Wearable Computers*, pp.153-154, 2009

[7] http://www.lozano-hemmer.com/projects.php.

[8] J. Freeman, "Large Audience Participation, Technology, and Orchestral Performance," in *2005 International Computer Music Conference*, International Computer Music Association, 2005, pp. 757-760.

[9] http://www.biopac.com.

[10] M. Stephen, W.V. Bradley, V. Sandrine, K.S. Bennett, and R. Rogger "Influences of Large-Scale Form on Continuous Ratings in Response to a Contemporary Piece in a Live Concert Setting," *Music Perception*, 22(2), pp. 297-350, 2004.

[11] http://www.instantheartrate.com.

[12] D. Maynes-Aminzade, R. Pausch, and S. Seitz, "Techniques for Interactive Audience Participation," in *Proceedings Fourth IEEE International Conference on Multimodal Interfaces (IEEE Comput. Soc)*, pp. 15-20, 2002.

[13] http://opensoundcontrol.org/spec-1_0.

[14] http://cycling74.com.

[15] http://www.openframeworks.cc/.

[16] http://www.syedrezaali.com/blog/?tag=addon.

[17] J. Hans, *Cymatics: A Study of Wave Phenomenon and Vibration*, Newmarket, NH: MACROmedia, 2001.

[18] C. Roads, *Microsound*, MIT Press, Cambridge, Massachusetts, 2001.

[19] http://www.lifeorange.com/.

[20] http://collider.co/art/yuan-yi-fan-emily-davis-gallery/.

[21] http://www.ccs.ucsb.edu/primavera/FightOrFlight.

[22] http://en.wikipedia.org/wiki/Fight-or-flight_response.

[23] C.W. Reynolds, "Flocks, Herds and Schools: A Distributed Model," *Computer Graphics*, 21(4):25-34, 1987.

[24] Y. Nagashima, "Bio-sensing System and Bio-feedback System for Interactive Media Arts," in *Proceedings of the International Conference on New Interfaces for Musical Expression*, pp 48-53, 2003.

A SONIFICATION OF KEPLER SPACE TELESCOPE STAR DATA

Riley Winton, Thomas M. Gable, Jonathan Schuett, & Bruce N. Walker

Sonification Lab, School of Psychology
Georgia Institute of Technology
654 Cherry Street Atlanta GA 30332 USA
{rjwinton,thomas.gable,jonathan.schuett,bruce.walker}@gatech.edu

ABSTRACT

A performing artist group interested in including a sonification of star data from NASA's Kepler space telescope in their next album release approached the Georgia Tech Sonification Lab for assistance in the process. The artists had few constraints for the authors other than wanting the end product to be true to the data, and a musically appealing "heavenly" sound. Several sonifications of the data were created using various techniques, each resulting in a different sounding representation of the Kepler data. The details of this process are discussed in this poster. Ultimately, the researchers were able to produce the desired sounds via sound synthesis, and the artists plan to incorporate them into their next album release.

1. INTRODUCTION

A representative of a professional group of musicians recently came to the authors for advice on how to properly sonify data obtained from stars. The initial request was for the researchers to produce any kind of musical result with the only constraint being that the sounds must be produced purely from star data. In other words, no artificial constructions or manipulations of sound were to be accepted. Ideally, the sonifications produced were to be utilized in the musicians' next major album release and therefore also required the sounds to be musical in some form. In an effort to create these desired sounds the project extended into the realms of many relevant subfields of sonification including digital signal processing, audio synthesis, and general sound design. Each of these approaches is presented in this poster.

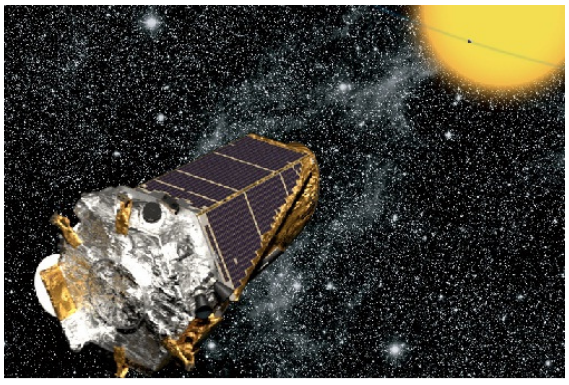


Figure 1: Artist's rendition of the Kepler spacecraft [1]

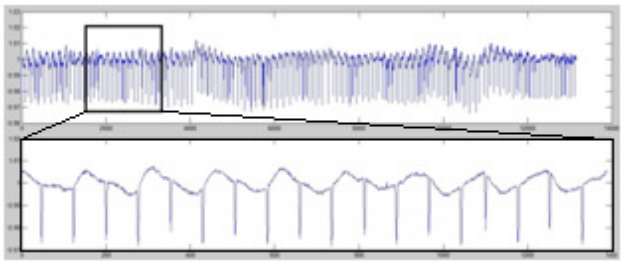


Figure 2: The top graph shows an original waveform as retrieved from planethunters.org, as graphed from within MATLAB. The black inset highlights one particular segment while the lower graph is a magnified image of that signal.

2. KEPLER DATA

The data that the authors were given were produced by NASA's Kepler space telescope pictured in Figure 1 and were gathered from the public site planethunters.org [2]. The data contained brightness values for certain stars across long periods of time. This data set was created by the telescope in the search for terrestrial planets within habitable zones of stars [1]. The fluctuations in the brightness values represent when a planet is passing between the Kepler telescope and the star it is focused on, an example of which is displayed in Figure 2.

3. SONIFICATION SANDBOX

Initially, the authors were asked to produce a general sonification for a sample set of data. To comply, the Sonification Sandbox [3] was utilized. This software package is available freely to the public, and was designed by researchers from the Georgia Tech Sonification Lab. For this first iteration, the data were simply imported into the software and the data values were mapped to various MIDI pitches. This technique was used as it is a standard first step when sonifying data. The Sonification Sandbox automatically handles mapping the values to certain pitches, which means the user is free to adjust further parameters such as timbre, tempo, frequency range, and others. An issue that arose with this approach involved the limited variability of the data sets. Since most of the stars' brightness values centered around 1 and the standard deviation was frequently on the thousandths scale, the software often had issues assigning frequency values to the clustered data points. To correct this, the data were standardized in a way that occupied a significantly larger range of values and was centered

on zero. An example of this output can be heard by listening to "0-Sandbox.midi."

4. AUDIFICATION IN MATLAB

The resulting samples from the Sonification Sandbox attempt were well received, however, the sound produced was not what the musicians wanted to use in their project. More specifically, the musicians were looking for timbres comprised of star data, not sequences of sonified musical pitches. In an effort to make the sounds more closely reflect the "heavenly" tones the musicians wanted, the authors tried a different approach. Instead of sonifying by mapping the brightness values to discrete MIDI pitches, the entire dataset was imported into MATLAB and audified (see [4]) via the soundsc() function. This simply plays back the data as a waveform while automatically scaling the values to an appropriate range. An example of this raw audification can be heard by listening to "1-MATLAB.wav." Since sample length was not a concern at this point in the research, sampling frequency was manipulated in order to quickly change pitch. This resulted in a much more applicable tone.

5. ADAPTING THE TONE

After this point, the musicians decided that the type of sound was beginning to approach the desired characteristics and all that needed to be refined was the timbre. The musicians were searching for a very specific kind of tone—one that could be described as heavenly or angelic. To solve this, the authors sifted through the numerous data sets located at planethunters.org in order to find a clean signal that would produce the desired timbre. Sinusoidal, periodic, or otherwise regularly patterned data sets were targeted and several were found. The previous procedure with MATLAB was then used to produce a new set of waveforms with star SPH10105467, as seen raw in Figure 3 and cleaned in Figure 4.

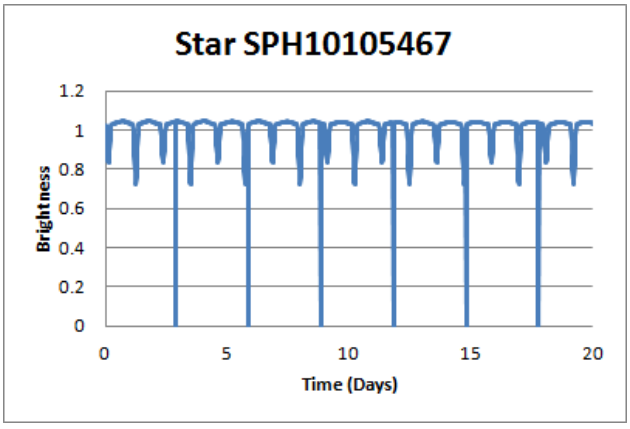


Figure 3: The graph shows the original waveform for star SPH10105467 as retrieved from planethunters.org. The vertical lines represent zeros (errors) in the data set. This is just one example of an artifact that contributed to noise in the original signal.

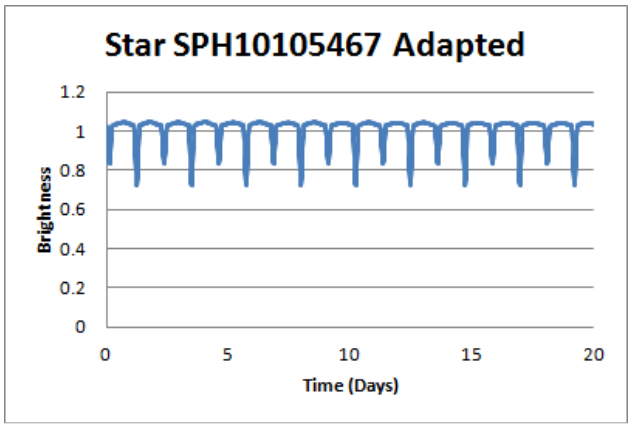


Figure 4: The graph shows the waveform for star SPH10105467 after cleaning the error values from the data set.

Further iterations of sounds were produced via experimentation with several different procedures. Upon initial creation of the waveforms in MATLAB, simple filtering was employed to clean the signals. For one series, as shown in Figure 2, one small segment of a signal was extracted. This small clip was then repeated numerous times, thus resulting in an extended version of this one segment. Simple bandpass filters were then applied in order to remove frequencies outside of the range 150 - 800Hz. This process yielded samples as heard in "2-Filtering.wav." This tone was much closer to the expected timbre, but it had some unnatural harmonics. The artists still desired a cleaner tone, so the researchers utilized the curve fitting tool in MATLAB in order to further remove some of the noise and artifacts that were evident in the natural data. This process yielded a more aesthetic tone, and it can be heard by listening to "3-CurveFitting.wav." The musicians greatly preferred the pure tone that this process yielded, and they then requested a set of 24 different musical pitches to compose the melody that will ultimately be used in the album. A short excerpt of this raw melody can be heard by listening to "4-Melody.wav."

6. PERFECTING THE MELODY

Once the previous waveforms were compiled into a short melody by the musicians, the authors decided to experiment with the sound's timbre further by applying an amplitude envelope. They found this to be the best route to maintain the clean pitch of the sounds and still make them have the desired timbral characteristics. This was accomplished by using data from another star (SPH10105611, see Figure 4) and applying those data points as an amplitude envelope onto the carrier signal of the composed melody. This method was accomplished by using MATLAB's interp1() function, linear interpolation, to stretch the envelope signal to the length of the melody. After scaling the envelope signal to a range between 0-1, the times() function was used to multiply the melody and envelope vectors together, thus applying a musical tremolo effect onto the entire melody. The resulting track was then sent to the musicians, who deemed it to be satisfactory for their final project. The sample melody excerpt from earlier can be heard with the tremolo effect by listening to "5-Envelope.wav."

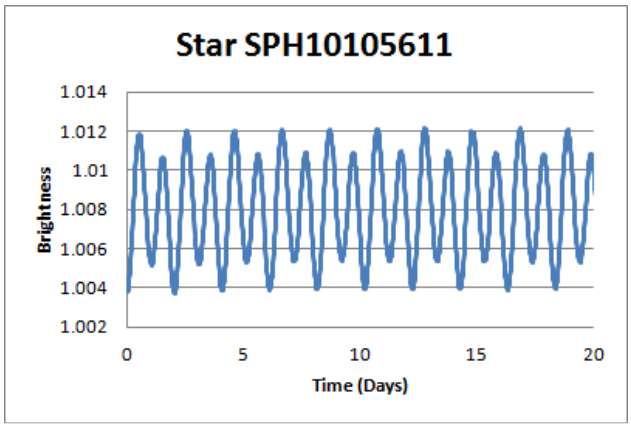


Figure 5: The graph shows the waveform for star SPH10105611, the star used as the amplitude envelope. This star was chosen because of its relatively even periodicity which translates well for a tremolo envelope.

7. CONCLUSION

The prior step marked the conclusion of this research endeavor. Currently, the music production is still in press with a release expected to be imminent within the next year. The musicians intend to incorporate a full orchestration atop the given melody, but this final version has yet to be released. Future works related to sonification and audification may yet provide some interesting results, as there are still many other avenues through which one could construct different versions of sonified star data. Regardless, this research yielded an authentic yet aesthetically satisfying auditory construction of star data, while still enabling the musicians to produce their composition with natural sounds made only from one quantitative property of a few stars. This work serves as a guide for future projects in sonification and audification; especially those that have an aesthetic or musical aspect to consider.

8. ACKNOWLEDGMENT

The authors would like to thank David N. Fowler from Echo Movement for allowing us to use these lessons to write a paper about them and for coming to us with such an interesting research problem.

9. REFERENCES

[1] Kepler: A search for habitable planets. (2012). Retrieved May 9, 2012, from: http://kepler.nasa.gov/

[2] Planet Hunters. Planet Hunters Candidates List [Data file]. Retrieved from: http://www.planethunters.org/candidate_list.csv

[3] Walker, B. N. (2010). Sonification Sandbox (Version 6.1) [Software]. Atlanta, GA: Georgia Institute of Technology. Retrieved from: sonify.psych.gatech.edu/research/sonification_sandbox/download.html

[4] Walker, B. N., & Nees, M. A. (2011). Theory of Sonification. In T. Hermann, A. Hunt, & J. Neuhoff (Eds.), *The Sonification Handbook* (pp. 9-39). Berlin, Germany:Logos Publishing House.

EVALUATION OF A MATLAB-BASED VIRTUAL AUDIO SIMULATOR WITH HRTF-SYNTHESIS AND HEADPHONE EQUALIZATION

György Wersényi

Széchenyi István University
Department of Telecommunications
H-9026, Győr, Hungary
wersenyi@sze.hu

ABSTRACT

Virtual audio simulators are the basic tools for localization experiments focusing on the parameters of the applied HRTFs. This paper presents a software solution capable of playback monotic wave files, setting the individual ITD information based on head-diameter data, HRTF filtering in 1-degree spatial resolution for static and dynamic sound sources and record output stereo wave files for headphone playback. Furthermore, a headphone equalization tool is implemented based on measured and averaged transfer functions. Based on this, generated IIR or FIR equalization filters can be applied on the sound data for scientific purposes.

1. INTRODUCTION

Virtual audio displays (VADs) try to simulate sound environments, where listeners are embedded or, at least, exposed to sound sources with directional information. This is usually achieved by setting the Interaural Time and Level Differences (ILD and ITD), extended by filtering of the outer ears, the Head-Related Transfer Functions (HRTFs) [1]. The virtual or augmented reality simulators may include visual and tactile interfaces as well. Regarding sound, the playback system should be able to perform digital filtering and signal processing methods to create binaural signals over headphones. Fig.1. shows the usual description of the transmission and required measurements based on *Wightman and Kistler* [2].

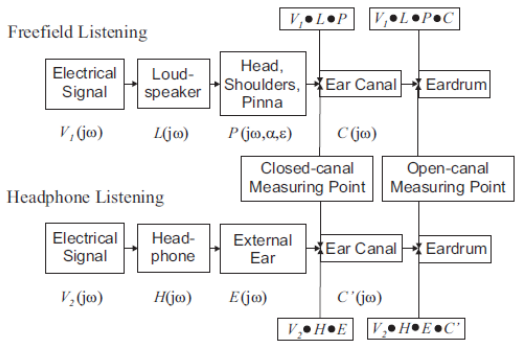


Figure 1. *Virtual audio simulation over headphones, compared to free-filed listening presented by Wightman and Kistler [2].*

Several parts and elements of this transmission can be investigated. The focus of experiments includes:

- overall quality, spatial resolution of applied HRTFs (number, length, frequency resolution), as well as the effect of interpolation of missing directions,
- the use of individual, non-individual (generic) and dummy-head HRTFs, that is, the effect of „quite similar looking" but different HRTFs (role of fine structure in the frequency),
- the effect of the minimal-phase reconstruction, which means HRTF filtering only by the magnitude characteristics followed by additional ITD settings, based on mathematical modeling (rigid sphere, head-diameter),
- the effect of headphone equalization including measurement and averaging,
- possible effects of dynamic, real-time adaptation based on head-tracking and room reverberations.

Our former research focused on the measureable effects in the spectral fine structure of dummy-head HRTFs. It revealed how the environment near the head influences the HRTFs [3-5]. HRTFs of the naked torso as well as HRTFs using clothing, glasses, hats, hair etc. were measured with high accuracy. Automated evaluation based on the spectral differences (HRTFDs) showed differences up to 5-15 dB from the same direction in the fine structure of the HRTFs indicating strong effects of the acoustic environment near the head. However, real-life localization usually does not suffer from decreased performance with and without glasses or before and after having a haircut. On the other side, virtual simulation is very sensitive to small changes in the HRTFs. Our goal is to test whether participants are able to recognize different customized HRTF sets on a VAD, based on the mentioned database. Virtual localization tests are planned with different HRTFs to determine the effect of everyday objects near the head. With other words: are subjects able to determine HRTFs with or without hair, or at least detect differences in their localization performance? Furthermore, the system has to be capable of evaluate other parameters such as head-tracking, headphone equalization and minimal-phase assumption.

Measurements and results of listening tests aiming at the effect of the acoustical environment with hair and clothing are rare in the literature, especially if high resolution and accuracy are needed. Measurements of *Tarnóczy, Riederer* and *Treeby* suggested effects similar to ours [6-18]. Note, that the use of a

dummy-head allows increased measurement accuracy, and using the method of spectral differences (HRTFDs) individual properties can be eliminated.

2. THE "3D SOUND GENERATOR" VIRTUAL AUDIO SIMULATOR

The main part of the setup is the virtual audio simulator. Mono wave files can be loaded, played back and analyzed (time and frequency domain plots). The dummy-head HRTF files can be loaded for the left and right ear respectively for a given angle (static source) or for dynamic simulation in the horizontal plane. In this case, a circling movement around the head is simulated in a user-determined spatial resolution. The best resolution is one-degree. The HRTFs are automatically plotted for visual feedback. After clicking „PROCESS!", the input file will be filtered by the HRTFs in the frequency domain. Using the minimal-phase assumption, only the HRTF magnitude is used for filtering. Finally, based on the head diameter data, the ITD information is calculated and added (HRTF customization). Currently the program uses the Woodworth-formula for this:

$$ITD \approx \frac{d(\varphi + \sin\varphi)\cos\delta}{2c} \quad , \qquad (1)$$

where d is head diameter, c is speed of sound, φ is azimuth, δ is elevation if applies [19, 20]. The resulting stereo wave file can be played back and saved. Currently, HRTFs of the naked torso from the horizontal plane are used, but they can be changed by replacing them with different sets in the dedicated subfolder (e.g. those with hair, glasses or with different elevation). The system works off-line: exported wave files can be played back without having MATLAB installed. The saved stereo wave files can serve as input files for the headphone equalization module. Fig.2. shows the screenshot of the GUI.

The main goal of the system is to test different HRTF sets and audible effects of the modified acoustical environment (hair, glasses, clothing) during virtual simulation. Furthermore, we would like to test different ITD approximation methods, whether there is any noticeable difference between the Woodworth-method and other formulas (to be implemented). Finally, the system is planned to be extended by head-tracking using a simple „smartphone" instead of having expensive motion-trackers such as the NEC/TOKIN 3D motion sensor or the Polhemus tracker. An Android-based program helps to get the gyroscope and accelerometer information of the cellphone and communicates with the laptop in real-time via the infrared or bluetooth connection. Spatial accuracy and latency has to be investigated.

3. THE HEADPHONE EQUALIZATION TOOL

Headphone equalization can be made using a separated MATLAB software. First, two headphones were measured in the anechoic chamber using the same Brüel & Kjaer dummy-head and individual measurements were made with the Brüel & Kjaer 4101 binaural microphone [21]. Left and right side were measured parallel but independently. Measurements were repeated ten times after re-placement of the headphone on the head, and results were stored in the PULSE LabShop format. These files, as well as other input formats (wave files, txt files etc.) can be used by the program for calculations.

The MATLAB application can import various data formats of other measurements if needed and calculates averaged complex transfer functions for left and right side respectively (Fig.3). Window-functions are also available as well as setting of the DC attenuation.

Table 1 and Figures 4-6 show the possibilities of creating filters in MATLAB. All three methods are implemented and can be selected by the user.

IIR filter		**FIR filter**	
Method	**Function**	**Method**	**Function**
Yule-Walker („Least-Square")	yulewalk	Sampling	fir2
Identify discrete-time Filter	invfreqz	Least-Square	firls, fircls
		Windowing	fir1
		Interpolation	intfilt
		P-normed	firlpnorm
		Nyquist-filter based	firnyquist

Table 1. Methods and functions in MATLAB for creating IIR and FIR filter coefficients.

YULEWALK creates an IIR filter without phase information. Order of nominator and denominator is the same. A low order filter is suitable for the estimation of the target magnitude and the filter is usually stable (zeros are inside the unit circle). However, calculations seemed to be depending on the hardware configuration.

INVFREQZ also creates an IIR filter but with phase information. Order of nominator and denominator is different. Although the filter is usually not stable, it can be used for emulation. Computation time is the largest.

Using the FIR2 method, only the nominator coefficients have to calculated, the denominator is one. It is always stable, has a larger order, but it is without phase information.

The program is able to create and save the filter coefficients, filter input wave files and check the stability of the filters based on the poles and zeros (Fig.7).

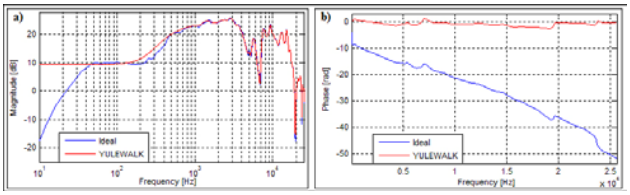


Figure 4. *YULEWALK method, 77-order estimation without phase information.*

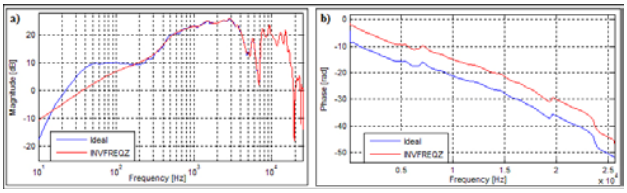


Figure 5. *INVFREQZ method, 64-order nominator and 96-order denominator estimation with phase information.*

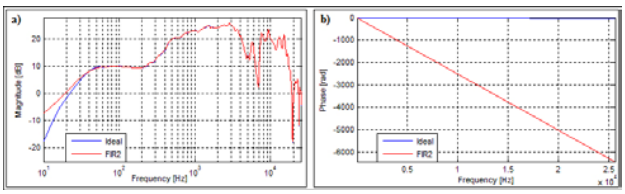


Figure 6. *FIR2 method, 4096-order estimation without phase information.*

4. CONCLUSIONS

This paper presented a MATLAB-based virtual audio simulator for scientific purposes. It includes real-time or off-line filtering of input data with the applied HRTF set. HRTFs originate from a dummy-head measurement system in different environmental conditions (naked dummy-head, clothing, glasses, hair etc.). The HRTFs can be customized by setting the ITD information based on the individual head-diameter and the Woodworth-formula. Output sound files can be exported in stereo for off-line applications. Static sources in 1-degree resolution as well as moving sound sources around the head can be emulated. Furthermore, the system can use several measured data formats of headphone transfer functions. Based on the mean transfer function, IIR and FIR filters can be generated as equalization filters for the playback. Future works includes listening tests with sighted and maybe blind participants primarily to test various HRTF sets and the audible effects and artifacts caused by the variations of the environment near the head; the effect of different ITD setting methods and the extension with head-tracking using a smartphone's gyro and accelerometer.

5. ACKNOWLEDGEMENT

The author would like to thank all the programmers and the participants in testing and debugging. The investigation and the publication was supported and funded by the Universitas-Győr Alapítvány in the framework „TáMOP 4.1.1./A-10/1/KONV-2010-0005" and by the "TáMOP 4.2.2/B-10/1-2010-0010".

6. REFERENCES

[1] H. Møller, "Fundamentals of binaural technology," *Applied Acoustics,* vol. 36, pp. 171-218, 1992.

[2] F. Wightman, and D. Kistler, "Measurement and validation of human HRTFs for use in hearing research," *Acta Acustica united with Acustica*, vol. 91, no. 3, pp. 429-439, 2005.

[3] Gy. Wersényi, and A. Illényi, "Differences in Dummy-Head HRTFs Caused by the Acoustical Environment Near the Head," *Electronic Journal of "Technical Acoustics" (EJTA)*, vol. 1, 15 pages, 2005. http://www.ejta.org

[4] A. Illényi, and Gy. Wersényi, "Evaluation of HRTF data using the Head-Related Transfer Function Differences*," in Proc. of the Forum Acusticum*, pp. 2475-2479, Budapest, 2005.

[5] A. Illényi, and Gy. Wersényi, "Environmental Influence on the fine Structure of Dummy-head HRTFs," *in Proc. of the Forum Acusticum*, pp. 2529-2534, Budapest, 2005.

[6] T. Tarnóczy, "Über den Verstärkerungs-Verminderungs-Effekt der Ohrmuschel und des Kopfes," in *Proc. of 6th Int. FASE Conference*, Zürich, Switzerland, 1992, pp. 229-232.

[7] P. F. Hoffmann, and H. Møller, "Some observations on sensitivity to HRTF magnitude," *J. Audio Eng. Soc*., vol. 56, no. 11, pp. 972-982, 2008.

[8] Gy. Wersényi, "Representations of HRTFs using MATLAB: 2D and 3D plots of accurate dummy-head measurements," *in Proc. of the 20th International Congress on Acoustics 2010 (ICA 2010),* Sydney, Australia, 2010, 9 pages.

[9] S. Carlile, and D. Pralong, "The location-dependent nature of perceptually salient features of the human head-related transfer functions," *J. Acoustical Soc. Am.***,** vol. 95, no. 6, pp. 3445-3459, June 1994.

[10] F. Asano, Y. Suzuki, and T. Sone, "Role of spectral cues in median plane localization," *J. Acoustical Soc. Am.***,** vol. 88, no. 1, pp. 159-168, July 1990.

[11] K. A. J. Riederer, *Head-related transfer function measuremnts,* Master Thesis, Helsinki University of Technology, 1998.

[12] K. A. J. Riederer, "Repeatability Analysis of Head-Related Transfer Function Measurements," *105th AES Convention Preprint 4846*, San Francisco, USA. 1998.

[13] K. A. J. Riederer, HRTF analysis: Objective and subjective evaluation of measured head-related transfer functions, Dissertation, Helsinki University of Technology, Espoo, 2005.

[14] B. E. Treeby, *The effect of hair on human sound localisation cues*, Dissertation, University of Western Australia, 2007.

[15] B. E. Treeby, J. Pan, and R. M. Paurobally, "The effect of hair on auditory localization cues," *J. Acoustical Soc. Am.***,** vol. 122, no. 6, pp. 3586-3597, December 2007.

[16] B. E. Treeby, R. M. Paurobally, J. Pan, "Decomposition of the HRTF from a sphere with neck and hair," in *Proc. of the 13th Int. Conf. on Auditory Display (ICAD 07)*, Montreal, Canada, 2007, pp. 79-84.

[17] B. E. Treeby, J. Pan, and R. M. Paurobally, "An experimental study of the acoustic impedance characteristics of human hair," *J. Acoustical Soc. Am.***,** vol. 122, no. 4, pp. 2107-2117, October 2007.

[18] B. E. Treeby, R. M. Paurobally, and J. Pan, "The effect of impedance on interaural azimuth cues derived from a spherical head model," *J. Acoustical Soc. Am.***,** vol. 121, no. 4, pp. 2217-2226, April 2007.

[19] P. Minnaar, J. Plogsties, S. K. Olesen, F. Christensen, and H. Møller, "The Interaural Time Difference in Binaural Synthesis," *108th AES Convention Preprint 5133*, Paris, France, February 2000.

[20] J. Nam, J. S. Abel, and J. O. Smith III, "A Method for Estimating Interaural Time Difference for Binaural Synthesis," *125th AES Convention Preprint 7612*, San Francisco, USA, October 2008.

[21] Gy. Wersényi, "On the measurement and evaluation of bass enhanced in-ear phones," *in Proc. of the 20th International Congress on Acoustics 2010 (ICA 2010),* Sydney, Australia, 2010, 6 pages.

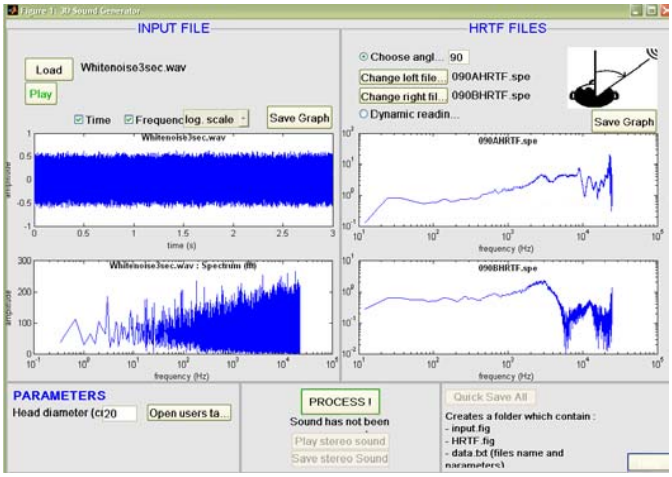


Figure 2. *Screenshot of the GUI of the 3D Sound Generator program. You can load an input file, set the head diameter, select the left and right HRTFs for static or dynamic emulation of sound source direction and you can save the result as a stereo wave file. Headphone equalization can be made with a separated program.*

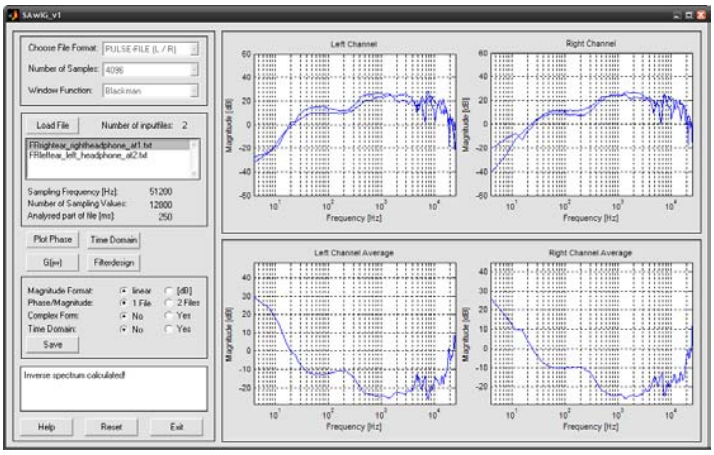


Figure 3. *Screenshot of the GUI. You can display measured results, calculate the average of repeated measurements (up to 10 left and right) both for magnitude and phase. The inverse is calculated by a complex 1/x in the frequency domain. It is also possible to save impulse responses corresponding to the measured or calculated data.*

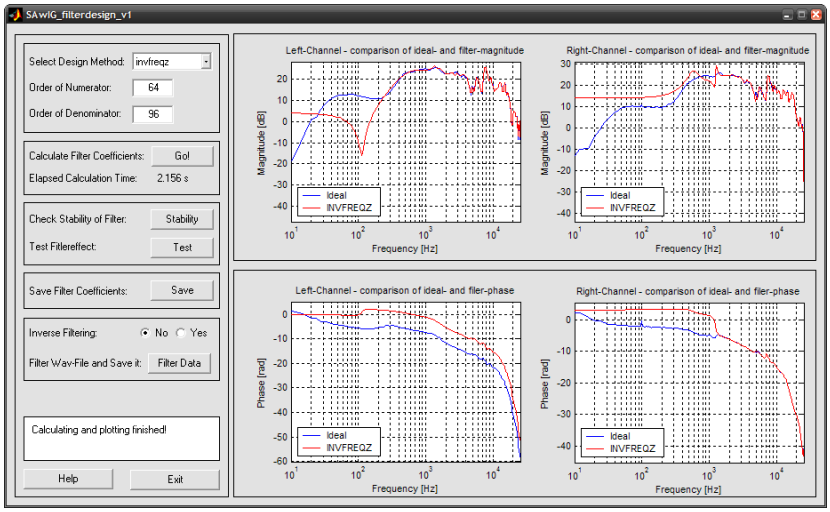


Figure 7. *Screenshot of the GUI of the headphone equalization tool, as part of the measurement module. Based on the inverted transfer function it is possible to create FIR or IIR filter coefficients directly, save them for further use and use them as filter for wave files.*

EEG SONIFICATION FOR EPILEPSY SURGERY: A CLINICAL WORK-IN-PROGRESS

Cole A Giller[1], Anthony M Murro[2], Yong Park[2], Suzanne Strickland[2], Joseph R Smith[1]

Departments of Neurosurgery[1] and Neurology[2]
Georgia Health Sciences University
Augusta, Georgia 30912, USA
{cgiller|amurro|sstrickland|ypark|jsmith}@georgiahealth.edu

ABSTRACT

Although EEG sonification has been well studied, its potential to identify seizure foci in patients undergoing epilepsy surgery has received little attention. We explore the sonification requirements needed for this application, and discuss a preliminary approach that has identified an auditory marker for epileptic tissue that agrees with standard EEG. Development of these early ideas into a clinical tool would be a welcome addition to the epilepsy surgery evaluation.

1. INTRODUCTION

As many as a third of patients with epilepsy will continue to have seizures despite optimal treatment with medication [1]. Because continued seizures can lead to neurologic decline and death, surgery may be offered to remove the portion of the brain in which the seizure begins (the 'seizure focus'). The decision for surgery is not trivial, and patient selection protocols are extensive. Foremost among these is the video-EEG, in which simultaneous real time video and EEG recordings are obtained continuously over several days in order to correlate the behavior of the patient during each seizure with the EEG. If the seizure type matches the EEG and other data, surgery can be offered.

In order for surgery to be effective, the seizure focus must be removed. Unfortunately, identifying seizure foci with EEG can be subtle, because they are heralded by low amplitude changes that are difficult to see and confounded by normal and abnormal rhythms as well as artifact. Failure to identify the focus is therefore not unusual, and can be discouraging for the epilepsy team and devastating to the patient for whom surgical options are blocked.

Figure 1: Grid of EEG contacts placed directly on the brain of one of our patients for seizure detection.

We describe a first approach to use EEG sonification to identify the seizure focus in EEG recordings obtained for evaluation for seizure surgery. Our hope is that the sophisticated auditory capabilities of the human ear can augment the more standard visual analysis of EEG data.

2. REQUIREMENTS OF EPILEPSY SURGERY

Specific demands of the surgical evaluation constrain our audification strategy in three ways. First, although the data can be analyzed offline, they must be viewed in real time to allow comparison with the video recordings. Second, it may be advantageous to analyze the EEG spectrum directly rather than use more sophisticated parameter mapping or event-based models [2,3]. This is because the goal is not to detect the seizures themselves – they are easily revealed by the EEG/video data – but rather, to detect EEG synchronies occurring shortly before the seizure that identify the first site of seizure onset. Because it is unclear which features of these synchronies are most important, and because it is precisely this data that sonification is to explore, we are reluctant to abandon the rich detail of the full spectrum. Furthermore, seizures considered for surgery are less stereotypical and less predictable than those of many other types of epilepsy, making the choice of parameters or models difficult. We therefore favor a more direct audification approach, at least at this early stage of our experience. Third, EEG data in the delta (<4 Hz), gamma (25 to 100 Hz) and ripple (100 to 500 Hz) ranges are valuable to epileptologists because activity in the first is an important indicator of epileptic tissue and activity in the others – especially the ripple range - is an important indicator of seizure onset [4]. This bandwidth is larger than has been previously considered [2,3] and poses special challenges to direct audification.

3. METHODS AND RESULTS

We present an early preliminary analysis of EEG data obtained from our patients being evaluated for seizure surgery.

3.1 Data and Conditioning

EEG recordings were obtained from electrodes surgically implanted in the brains of 5 patients being treated and operated upon by us for intractable epilepsy. Data from implanted electrodes were chosen rather than from scalp electrodes because they are relatively free from artifact, sample a more specific volume of brain tissue, and have known location relative to the proven seizure foci. Each recording contained data from the few minutes prior to a seizure, the seizure itself,

and a few minutes afterwards. The data was bandwidth filtered (time constant 2.0 seconds, low pass 70 Hz), sampled at 200 Hz and passed to the Matlab environment. The filter threshold of 70 Hz used for clinical work prevented examination of the gamma and ripple bands. Channels obtained from sites that later proved to be seizure foci were compared with channels obtained from sites without initial seizure activity.

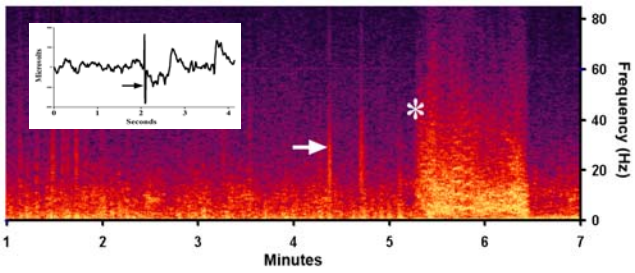


Figure 2: Time-frequency plot of EEG from implanted electrode in patient with seizures. Arrow indicates a 'shimmer'. Asterisk indicates start of seizure. Inset shows spike in raw EEG (arrow) corresponding to same 'shimmer'. Note difference in time scale.

3.2 Audification and Data Analysis

Vocoder software [5] was used to slow the timebase of each recording without changing the pitch of the EEG frequencies. An audio file was written from these slowed recordings using a high sampling rate to increase the EEG frequencies into the audible range. If f is the original sampling rate and 1/r is the factor by which the vocoder changes the timebase, then increasing the sampling rate from f to fr produces a recording in real time with frequencies altered by a factor of r. We chose to use r = 200 to shift the delta-to-gamma range of 1 to 70 Hz to the audible range of 200 to 14000 Hz.

The recordings were examined with software (Soundbooth, Adobe, San Jose, California) allowing visual examination of the waveform, its time-frequency spectrum, and auditory review of the recording. Frequency ranges of audible tones were correlated to bands visible on the time-frequency spectrum and to the raw EEG. Auditory review included the entire spectrum as well as selected isolated frequency bands.

3.3 Results

Artifact arising from the vocoder algorithm could be minimized by using short windows (512 points, 2.6 seconds) and small overlaps. Short segments of high pitched, high amplitude signals (1 to 2 seconds, 1 to 4 kHz) that could be described as acoustic 'shimmers' were frequent in tissue that eventually developed seizures, but were rare in normal tissue. Review of the EEG showed that they arise from 'spikes' in the EEG waveform that are classical makers of epileptogenicity. The well-known 1/f character of the EEG was apparent in the audio signals [6], with lower frequencies having higher amplitude than higher frequencies. In some cases, widening the lowpass filter allowed demonstration of frequencies in the ripple range as indentations superimposed on the 1/f shape of the EEG spectrum.

4. DISCUSSION AND FUTURE DIRECTIONS

The specific demands of using EEG sonification to find seizure foci in patients being evaluated for epilepsy surgery – the need for real time review, the uncertainty of key parameters, and the wide range of required frequencies – impose constraints on the sonification. Barriers include vocoder artifacts and the lack of classification of the auditory EEG data. Our early data identify audible features of the EEG ('shimmers') corresponding to classical markers of epileptogenicity ('spikes').

Perhaps the most important task is to determine if sonification can augment the already sophisticated visual EEG review. Plans are underway to establish workstations at our institution allowing auditory review of EEG data during the routine review of the clinical data. After gaining experience correlating these data, the hypothesis that auditory EEG can improve foci detection will be tested with a randomized trial using existing data for which the location of the seizure foci is known. At the same time, various listening strategies and technical refinements (e.g., methods such as differentiation to alter the 1/f shape of the EEG spectrum to better hear frequencies in the ripple range, algorithms to reduce vocoder artifact, and audification methods allowing auditory display of the entire range of epilepsy frequencies of 1 to 500 Hz) will be explored. Collaboration with established sonification groups will be actively sought.

5. CONCLUSION

We explore EEG sonification as an aid to the identification of seizure foci for epilepsy surgery. Development of this early experience into a clinical tool would be a welcome addition to those making these difficult decisions.

6. REFERENCES

[1] P. Kwan and M.J. Brody, "Definition of refractory epilepsy: defining the indefinable?," *Lancet Neurology*, vol. 9, no. 1, pp. 27-29, 2010.

[2] T. Hermann, P. Meinicke, H. Bekel, H.Rigger, H.M. Muller, S. Weiss, "Sonifications for EEG Data Analysis," in *Proc. Int. Conf. on Auditory Display (ICAD 2004),* Sydney, Australia, July 2004.

[3] G. Baier T. Hermann, U. Stephani, "Event-based sonification of EEG rhythms in real time," *Clin Neurophys*, vol. 118, pp. 1377-1386, 2007.

[4] J. Engel Jr, A. Bragin, R. Staba, I. Mody, "High-frequency oscillations: what is normal and what is not?," *Epilepsia*, vol. 50, no. 4, pp. 598-604, 2009.

[5] http://www.ee.columbia.edu/~dpwe/resources/matlab/pvoc

[6] C. Demanuele, C.J. James, E.J.S. Sonuga-Barke, "Distinguishing low frequency oscillations within the 1/f spectral behavior of electromagnetic brain signals," *Behavioral and Brain Functions*, vol. 3, no. 62, 2007.

NEW DIRECTIONS FOR SONIFICATION OF EXPRESSIVE MOVEMENT IN MUSIC

R. Michael Winters & Marcelo M. Wanderley

Input Devices and Music Interaction Laboratory (IDMIL)
McGill University CIRMMT
550 Sherbrooke St. West, H3A 1E3 Montreal, QC, Canada
Contact: Raymond.Winters@mail.mcgill.ca

ABSTRACT

Expert musical performance is rich with movements that facilitate performance accuracy and expressive communication. As in sports or rehabilitation, these movements can be sonified for analysis or to provide realtime feedback to the performer. Expressive movement is different however in that movements are not strictly goal-oriented and highly idiosyncratic. Drawing upon insights from the literature, this paper argues that for expressive movement in music, sonifications should be evaluated based upon their capacity to convey information that is relevant to visual perception and the relationship of movement, performer and music. Benefits of the synchronous presentation of sonification and music are identified, and examples of this display type are provided.

1. SONIFICATION OF EXPRESSIVE MOVEMENT

Recent developments in auditory display have infused human motion with sound for the purpose of analysis, motor learning, and adapted physical activity [1]. However, human motion is not limited to goal oriented movements like those frequently found in sports. In music for example, *expressive* [2] or *ancillary* [3, 4] gestures refer to movements that are not responsible for sound production, but nevertheless common in performance. Though complex and diverse – varying with the instrument, performer, and musical piece – these movements are otherwise highly consistent over time and reflect musical structure and expressive intention [5].

The use of high-resolution motion capture systems has enabled the quantitative study of these movements. In a typical setting, a performer wears reflective markers that are tracked overtime in three spatial dimensions using an array of calibrated infrared cameras. Due to the size and complexity of the data sets, sonification can be used to quickly browse through the data, make non-obvious relationships more apparent, and facilitate the process of data analysis.

1.1. Previous Work

The use of sonification for studying expressive gesture in performance began with a study of four clarinetists [6] who were asked to play the same piece of music with exaggerated, normal, and immobilized playing modes. Though mapping choices were discernible and could be used to expose data relationships that were not visually obvious, the mapping was not easily extendible to other performers due to the high variability in the movement patterns between subjects.

A more recent work [7] has compared Principle Component Analysis (PCA) and velocity of markers as preprocessing steps for sonification in a bimodal context using a "stickman" visualization. Using an open task, they found that sonification would work well in directing the attention of the user to aspects of the visual display in the velocity based mapping, but not in the PCA.

2. A NEW METHODOLOGY

Gesture in music performance is a rich field for sonification, but the expressive nature of these movements warrants special consideration that is distinct from goal-oriented movements that are common in sports. What is more important than the exact positions or velocities of points and angles on the body are the "higher-level" structural and emotional information they carry. This information can be organized around the relationship of movement performer and music, and what the movements convey to the viewer.

2.1. The relationship of movement, performer, and music

Building upon a foundational work in the study of expressive movement [4], there are three levels of gestures that need to be conveyed in sonification, the *material, structural,* and *interpretive*. Material gestures are those that are defined by the instrument being played. For example, the cello is more limited in possible expressive movements than the clarinet, resulting in different movement patterns. For a good sonification, a listener should be able to identify this type of difference.

The structural level of gesture concerns the relationship to the underlying music. For instance, highly difficult passages of music often impede mobility while easy passages and phrase boundaries see an increase in movement [8]. Though each performer moves differently, these sorts of structural cues are important and should be clear in sonification.

Finally, the interpretive gestures concern the performer's unique interpretation of the piece and convey their structural and emotional representation. For a good sonification, a listener should be able to identify two "takes" of the same performer playing a piece of music and likewise perceive that a different performer has played.

2.2. The perception of movement in musical performance

In the perception of music, the visual context provides cues that can modulate the emotional and structural perception of a piece. For instance, simply viewing a performer can extend the perceived length of phrases and reduce or augment ratings of tension [8]. In another study, [9] showed that the visual perception of regularity, fluency, speed, and amount of motion could predict the emotional ratings of happiness, sadness, and anger.

Results of [9] supported a possible invariance between viewing conditions, instrument, and musician. This invariance was supported by [10], who modificatied stickman avatars derived from motion capture data of real performers. Completely immobilizing the arms or torso, or even playing the avatar in reverse did not significantly effect judgements of tension, intensity, fluency, or professionalism. Increasing the amplitude of motion of the whole body was important however, implying this factor was more important than the movement of individual body regions.

If factors such as amplitude of motion are indeed more important to visual perception than the exact part of the body being moved, than it is wise that sonification of performers prioritize this cue. Additionally, if the regularity, fluency, and speed are important cues for conveyed emotion, likewise sonifications should focus on the ability to correctly display this information.

3. SONIFICATION FOR MUSIC-DATA ANALYSIS

New music research abounds with large, complex, time-varying data sets. For this data, sonification as a tool for analysis or display benefits from the shared medium of music and sonification. For gesture in particular, some of these benefits have already been identified by researchers using interactive sonification to teach bowing technique of the violin.

The first benefit, identified by [11], stressed that the shared temporal nature of music and the data could be used to understand data events as they occur temporally relative to the music. Later, [12] identified that for sonification and music research, listening is a familiar and widely used medium. Also, the shared acoustic medium could provide a more direct access to relationship of data and performance audio. For expressive gesture, this may provide a fuller display of the performer's expressive intension than the music alone, and may be closer to the performer's internal representation of the structural and emotional content of the piece.

A benefit that has not yet been identified is that through sonification, the visual aspect of musical performance is made accessible to the blind (or those who cannot see). If a sonification design is able to convey the structral and emotional cues discussed in Section 2, then it is a display medium that can be used to make expressive gesture accessible through sound.

Videos hosted on the IDMIL website[1] and Vimeo[2] provide examples of this display type. In the first example, a performer's expressive gestures are sonified and presented with performance audio and video. In the second example, sonification of the "eigenmodes" of a subject dancing to music [13] displays four metrical layers that can be compared to the metrical layers of the music itself. In both of these examples, sonification provides a dynamic display that conveys non-obvious information as well as the performer's unique representation of the piece.

4. CONCLUSIONS AND FUTURE WORK

This article has argued that for sonification, expressive movement should be treated differently than goal-oriented movement. Evaluation should be based upon the ability to convey movement cues that are relevant to visual perception and that highlight the relationship of instrument, music, and performer. Pairing music and sonification has benefits for analysis and display that are unique to their shared medium. In this way, a successful sonification can make expressive gesture accessible and provides a more complete display of a performer's expressive intentions in the same medium as the performed music.

[1] www.idmil.org/projects/sonification_project

[2] www.vimeo.com/peto/videos

5. REFERENCES

[1] O. Höner, A. Hunt, S. Pauletto, N. Röber, T. Hermann, , and A. O. Effenberg, *Aiding movement with sonification in "exercise, play and sport"*. Berlin, Germany: Logos Publishing House, 2011, ch. 21, pp. 525–553, o. Höner (chapter ed.).

[2] F. Delalande, *Élements pour une Sémiologie du Geste Musical*. Québec: Louise Courteau, 1988, pp. 85–111.

[3] M. M. Wanderley, "Non-obvious performer gestures in instrumental music," in *Gesture-Based Communication in Human-Computer Interaction*, A. Braffort, R. Gherbi, S. Gibet, J. Richardson, and D. Teil, Eds. Berlin: Springer Verlag, 1999, pp. 37–48.

[4] ——, "Quantitative analysis of non-obvious performer gestures," in *Gesture and Sign Language in Human-Computer Interaction*, I. Wachsmuth and T. Sowa, Eds. Berlin: Springer-Verlag, 2002, pp. 241–253.

[5] M. M. Wanderley, B. W. Vines, N. Middleton, C. McKay, and W. Hatch, "The musical significance of clarinetists' ancillary gestures: An exploration of the field," *Journal of New Music Research*, vol. 34, no. 1, pp. 97–113, 2005.

[6] V. Verfaille, O. Quek, and M. M. Wanderley, "Sonification of musicians' ancillary gestures," in *Proceedings of the International Conference on Auditory Display*, London, UK, 2006, pp. 194–197.

[7] F. Grond, T. Hermann, V. Verfaille, and M. Wanderley, "Towards methods for effective ancilliary gesture sonification of clarinetists," in *Proceedings of the International Gesture Workshop*, S. Kopp and I. Wachsmuth, Eds. Berlin, Heidelberg: Springer Verlag, 2009.

[8] B. W. Vines, C. L. Krumhansl, M. M. Wanderley, and D. J. Levitin, "Cross-modal interactions in the perception of musical performance." *Cognition*, vol. 101, no. 1, pp. 80–113, 2006.

[9] S. Dahl and A. Friberg, "Visual perception of expressiveness in musicians' body movements," *Music Perception*, vol. 24, no. 5, pp. 433–454, 2007.

[10] M. Nusseck and M. M. Wanderley, "Music and motion: How music-related ancillary body movements contribute to the experience of music," *Music Perception*, vol. 26, no. 4, pp. 335–353, 2009.

[11] O. Larkin, T. Koerselman, B. Ong, and K. Ng, "Sonification of bowing features for string instrument training," in *Proceedings of the International Conference on Auditory Display*, 2008.

[12] T. Grosshauser and T. Hermann, "The sonified music stand - an interactive sonification system for musicians," in *Proceedings of the Sound and Music Computing Conference*, 2009, pp. 233–238.

[13] P. Toiviainen, G. Luck, and M. R. Thompson, "Embodied meter: Hierarchical eigenmodes in movement-induced movement," *Music Perception*, vol. 28, no. 1, pp. 59–70, 2010.

SYSSON - A SYSTEMATIC PROCEDURE TO DEVELOP SONIFICATIONS

Katharina Vogt, Visda Goudarzi, Robert Höldrich

Institute for Electronic Music and Acoustics
University of Music and Performing Arts Graz
Inffeldg. 10/3, 8010 Graz, Austria
vogt@iem.at, hoeldrich@iem.at, goudarzi@iem.at

ABSTRACT

The newly started research project SysSon (http://sysson.kug.ac.at) will develop a systematic procedure to develop sonifications, and test the procedure with climate data. The SysSon approach addresses the relevant obstacles that are met when introducing sonification in a new scientific domain: the cultural bias, usability and technical issues. This paper presents the research approach and puts it up for discussion.

1. INTRODUCTION

Usual obstacles to the application of sonification in science have been cited, e.g. [1]. These include, amongst others, 1) a cultural bias, i.e. a listening comprehension barrier, as there are few traditions of using sound to do science and practically no training in it; 2) quality control and questions of usability; 3) working premises, i.e. a technical barrier, e.g., created by the fact that audio software is not compatible with data in the domain sciences. In SysSon, we want to address all these factors explicitly:

- The cultural bias is usually the strongest barrier. Therefore we will adjust the sound design explicitly to cultural metaphors of the domain science. Furthermore, a common terminology will be built up accompanying a sound library, which shall allow communicating about the sounds. The additional gain of the sonification approach will be pointed out by comparing it with advanced visualization in the domain.
- Quality control and usability need to be assured by vigorous evaluation. The project includes therefore several evaluation steps of the sonification design within different test groups. Furthermore, a public media installation and an expert workshop will be used to evaluate the project's results.
- The technical barrier can be treated by providing an independent, easy-to-use sonification tool at the end of the project, which is adjusted to software and data formats that are common in the domain science.

We will develop a systematic procedure taking these factors into account and elaborate sonifications for complex, dynamic data, as can be found in various fields. In the project, we chose climate data as an ideal case study. Climate data provide a good, practicable working basis, as both model data and measurement data are at hand, and they provide a straightforward real-world interpretation. The data sets are high dimensional and large. Furthermore, there is consensus on global climate change and the necessity of intensified climate research today in the scientific community and general public.

2. RESEARCH APPROACH

The research approach is based on our experience from previous projects (www.sonenvir.at, www.qcd-audio.at) and on a variety of knowledge of the ICAD community, of which not all projects can be cited here. SysSon is the systematic development and evaluation of a sonification design for the example case of climate data. It proposes a procedure for developing sonifications that are well integrated into the specific scientific community. The systematic sonification procedure of SysSon encompasses several steps:

Preparatory steps: As preparatory steps, the data has to be prepared, and a short update of the literature survey on current sonification strategies in the domain science field has to be conducted. Furthermore, the needs of the domain scientists have to be analyzed and existing visualization tools assessed according to their capabilities.

- Data preparation and literature survey
- Evaluation of existing visulization tools
- Analysis of domain scientists' needs

Interdisciplinary communication: In a second step, the interdisciplinary communication has to be built up between the specific language and metaphors of the domain scientists, and the one of the sonification designers. An extended TaDa (Task and Data analysis [2]) can be used for this part of the procedure. The metaphoric sonification methods [3] will be used to explore (implicit and explicit) audio and other metaphors of the domain scientists. With this knowledge, and based on our experience, a first library of sounds shall be established, which serves as a working basis for the sonification design. Once a sonification design has been developed (based on the evaluation cycles as described below), a final sound library and terminology can be assembled. The library serves as a key for the sonification (in analogy to the key of a graph), and facilitates a joint terminology of domain scientists and sonification designers. Sound phenomena in the sonification can be verbally described, understood, and, thus, better recognized and discussed.

- Analysis of domain scientists' metaphors
- Establishment of sound library

Sonification Design: The development of the sonification design is an iterative process based on the study of the domain metaphors. It comprises the choice of (a/) basic sonification method/s, the possibilities of user interaction, and the set of parameters, which are adjusted to the data.

- Development of sonification model
- Implementation of sonification

Evaluation: The sonification design is driven by a cyclic evaluation process. We propose three different test groups; the domain scientists as experts on the one hand, and non-experts, but aesthetically trained people - musicologists and sound engineers/ computer musicians - on the other hand. The domain scientists can use the sonification prototypes for exploration tasks and evaluate the scientific gain of the representation. The second and third group are responsible for an aesthetic evaluation, assuring that the sounds will not become annoying even when working long time with them. This group will also conduct simple exploration tasks. Open floor is given to a general public, who will give indirect feedback on the sonification in a media installation. Finally, sonification experts will discuss the project's theoretical outcome and the specific sonification design in a concluding workshop.

- Cyclic sound evaluation by three test groups
- Public and expert evaluation

Dissemination: Sound shall be used as a new means to display scientific data, but as an innovative medium also further spread the information to a general public, e.g., in a media installation. As deliverable, the sonification design has to be brought to a profound technical shape, which can be easily used by the domain scientists to work with.

3. CASE STUDY

A systematic sonification approach cannot be developed per se, but needs a meaningful case study of data. We chose climate data which is provided by the Wegener Center for Climate and Global Change (WegC, www.wegcenter.at). First results of the preparatory steps are shortly presented below.

3.1. Evaluation of existing visualization tools

Many challenges of sonification of a given data set are also found in visualization - the innovative data exploration software SimVis [4] (www.simvis.at, see Fig.1) has partly been developed in cooperation with the WegC. Other software used at WegC include IDL (Interactive Data Language, www.exelisvis.com) and the open source language R. We assess the use and functionality of these and other tools during the initial user interviews (see below).

3.2. Analysis of climate scientists' needs and metaphors

Eighteen members of the staff of WegC have been interviewed by a moderator and observed by a second interviewer, and were audio recorded. General questions assessed their qualification, the use and usability of their tools of data analysis and visualization, their user goals with these tools (incl. task success, efficiency, and learnability), and expectations for the sonification. The central part of the interview consisted of a self-chosen typical task that they walked us through, starting from raw data, where they have recently gained some insight during an analysis process. In a second step, the results from this task were presented in a 'group meeting', allowing us to study the communication in-between an expert focus group, thus assessing the typical language and metaphors of the field.

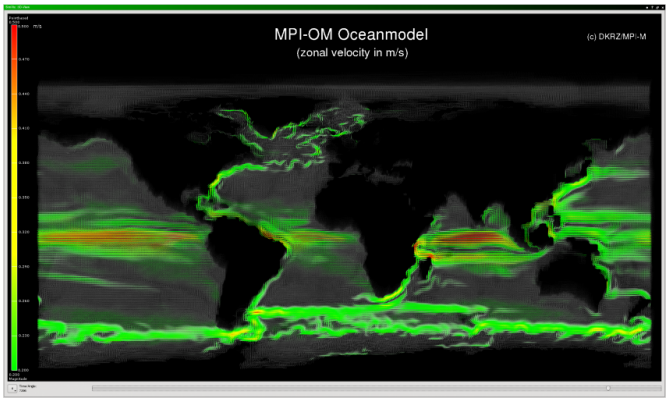


Figure 1: Visual analysis of ocean currents using SimVis, http://www.simvis.at/references/showcase (accessed 16/02/2012).

The analysis of the interviews is ongoing. It will comprise (a) a listing and research of data analysis tools, (b) the inquiry of a typical workflow in climate data analysis, (c) a language investigation of the transcribed interviews and focus group meetings. Furthermore, the acceptance of learning and using a new tool is quested, which shall also be indirectly raised by engaging the researchers from the beginning into the design process. The participants were rewarded with headphones to thank for their collaboration, and to further engage them with audio, e.g., by regularly sending them links of sound material resulting from the project as disseminated at http://soundcloud.com/syssonproject.

4. CONCLUSION

This extended abstract gives a quick overview over the planned systematics that will be further developed and tested in the research project SysSon, which has started in February 2012. Due to the shortness of this format, we cannot go into details with the planned sonification design and technical implementations, but rather want to stimulate a debate on the suggested research approach and the test design of the preparatory tests at WegC.

5. REFERENCES

[1] K. Vogt, "Sonification of simulations in computational physics," Ph.D. dissertation, University of Music and Performing Arts Graz, 2010.

[2] S. Barrass, "Auditory information design," Ph.D. dissertation, The Australian National University, 1997.

[3] K. Vogt and R. Höldrich, "A metaphoric sonification method - towards the acoustic standard model of particle physic," in *to be published in Proc. of the International Conference on Auditory Display*, 2010.

[4] F. Ladstaedter, A. Steiner, B. Lackner, G. Kirchengast, P. Muigg, J. Kehrer, and H. Doleisch, "Simvis: An interactive visual field exploration tool applied to climate research," in *New Horizons in Occultation Research: Studies in Atmosphere and Climate*, A. K. Steiner, B. Pirscher, U. Foelsche, and G. Kirchengast, Eds. Springer Berlin Heidelberg, 2009, pp. 233–244.

WHO'S SONIFYING DATA AND HOW ARE THEY DOING IT? A COMPARISON OF ICAD AND OTHER VENUES SINCE 2009

Nick Bearman

Associate Research Fellow in GIS
European Centre for Environment and Human Health
nick.bearman@pcmd.ac.uk

Ethan Brown

Statistical Consultant
statisfactions.com
ethancbrown@gmail.com

1. INTRODUCTION

What disciplines are applying data sonification, and what synthesis tools are they using to make the sounds? These questions are basic to understanding the state of sonification today, but they are surprisingly difficult to answer. This short review attempts to fill this gap by distilling common patterns of data sonification research. We hope that this will complement other literature reviews and give potential and current sonification researchers a sense of what is happening in the ICAD community, where there is room for new ventures, and where there is already a lot of active research to connect with. Additionally, we place ICAD in context with other academic publications.

Over its twenty years, ICAD participants have presented a wide variety applications for data sonification. Other reviews of the literature have already given general overviews of the work in the field [1], looked at how various physical quantities have been sonified [2], and how they were evaluated [3]. Instead, we wanted to focus on the people doing sonifications to get a current sense of which disciplines are involved in applied sonification and what tools they use.

The review covered 51 articles (29 in ICAD, 22 elsewhere) applying data sonification since 2009. Some ongoing studies have several published articles associated with them; however, we analyze all papers separately. The criteria for inclusion were whether a sonification example was created in the work (as opposed to a theoretical discussion or general presentation of a software tool) and whether they used data in the example sonification. The data could be real-world data or synthesized. A full list of the papers included in the review is available at `http://www.zotero.org/groups/icad_2012_sonification_tools/items`.

2. COLLABORATION AND SOFTWARE AT ICAD

Applied data sonification articles at ICAD were almost always affiliated with a music or technology department. The first authors on 22 of the 29 articles had a music/technology affiliation, and three more papers had a music/technology affiliation further down the author list. Institutions associated with the applied subject area–i.e. the source of the data being sonified–were not as prevalent, but did have a narrow majority (17 papers). Twelve articles involved a collaboration between a music or technology department and department in the subject area. Physics and biology were both well-represented in the applications, but there was no social science applications besides for one economics-related article [4], despite the fact that the social sciences are rife with quantitative data.

The prevalence of music and technology specialists in the literature is hardly surprising–sonification today invites that level of specialized knowledge to actually realize the complex sounds involved. To ease the use of sonification to explore data, several software toolkits have been created (e.g. the Sonification Sandbox, SoniPy, AesSon, and the Interactive Sonification Toolkit). Yet only one recent ICAD paper used a general sonification tool, and this paper was written by the author of the tool: David Worrall used his SoniPy framework to sonify capital trading data [4]. This echoes the frequent lament that there are no mature general-purpose data sonification toolkits [5].

Almost all of the ICAD papers used open-source computer music synthesis software to realize the sonifications (see Figure 1). SuperCollider was the most popular, accounting for 9 of the 29 papers; Pure Data, another open-source synthesizer, was almost as popular (7 papers). Csound and ChucK were rarely used, and the proprietary Max/MSP was used twice. There were no ICAD papers which used built-in MIDI software synthesis, which is one of the easiest ways to generate sound (many computers and mobile devices come with a MIDI software synthesizer). The remaining papers used a smattering of custom hardware and software for creating the sonification.

3. ICAD'S DATA SONIFICATIONS COMPARED WITH OTHER VENUES

The 22 non-ICAD papers we found had a lot of overlap in content and authorship with the ICAD community (although see *Limitations* section below). Only four of the non-ICAD papers we found had authors who had not previously appeared in ICAD; seven had authors who had all appeared in ICAD, and 11 had a mixture. However, only five articles were collaborations between music/technology departments and an institution in the applied data field. There were no sonifications that related specifically to social science data.

Pure Data was, again, a popular synthesis tool among the non-ICAD group, accounting for 5 of the 22 papers (see Figure 1). Unlike in the ICAD articles, SuperCollider was only used in 3 papers, and solutions using built-in MIDI software synthesizers were the most popular (6 papers).

In the full pool of 51 data sonification articles since 2009, authors with multiple recent publications tended to use the same tools. This suggests that the technical ease of using familiar software may override the advantages of alternate tools for different applications. Among the 22 authors who appeared on more than one publication, only 5 authors used more than one tool. For the 9

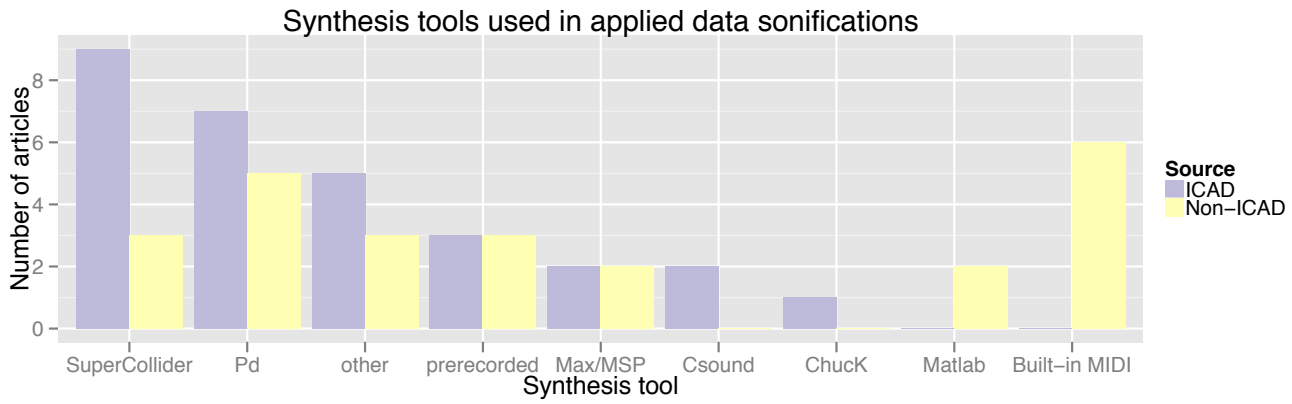


Figure 1: SuperCollider was easily the most popular synthesis tool among applied ICAD sonifications, but Pure Data (Pd) and built-in MIDI software synths were most common in non-ICAD articles.

authors who appeared first on multiple publications, there was only one exception: Nina Schaffert, who generally used Pure Data for sonification, used custom synthesis hardware in an early iteration of her rowing sonification system [6].

4. LIMITATIONS

A survey of sonification practitioners may be more effective than a literature review as a way of understanding the what, why, and how of sonification research today. It would allow us to ask people why they were conducting the research, what their original aims were, and why they used the tools they did. We initially started our review looking at other aspects of the sonifications, including the context, purpose, type of data, details of the user evaluation (if any), and the target user group. However, these were quite complex to define or were simply not well-described in the papers.

We believe that there are many other articles on sonification besides the ones we were able to find by searching for the keyword "sonification" on Google Scholar and Web of Science. The SAS Institute (a leading statistical software vendor) recently published research on auditory graphing without a single mention of the term "sonification" [7].

5. CONCLUSION

Sonification has not yet found its scientific champion. In Quetelet and other 19th-century innovators, visualization found leaders in applied fields such as economics who could also effectively promote new means of communicating and discovering their findings [8]. Also, several quantitative fields have very little representation in the sonification community, especially the social sciences. Existing sonification-specific tools are not gathering enough of a user-base beyond their authors to encourage the development of a mature piece of software. Instead, data sonification is proceeding with an interdisciplinary approach, often via collaborations between applied researchers and those with the technical and artistic expertise to use their favorite computer synthesis tool in order to realize the sonifications.

6. ACKNOWLEDGMENT

The authors would like to thank Benjamin Davison of Georgia Tech for his participation in the early phases of this research, and the two anonymous reviewers for their helpful suggestions.

7. REFERENCES

[1] D. Worrall, "Chapter 2: An overview of sonification," in *Sonification and information: Concepts, instruments and techniques (PhD dissertation)*. Canberra, Australia: University of Canberra, Mar. 2009.

[2] G. Dubus and R. Bresin, "Sonification of physical quantities throughout history: a meta-study of previous mapping strategies," in *International Conference of Auditory Display*, Budapest, Hungary, June 2011.

[3] K. Vogt, "A quantitative evaluation approach to sonifications," in *International Conference on Auditory Display*, Budapest, Hungary, June 2011.

[4] D. Worrall, "The use of sonic articulation in identifying correlation in capital market trading data," in *International Conference on Auditory Display*, Copenhagen, Denmark, June 2009.

[5] J. H. Flowers, "Thirteen years of reflection on auditory graphing: Promises, pitfalls, and potential new directions," in *International Conference on Auditory Display*, Limerick, Ireland, July 2005.

[6] N. Schaffert, K. Mattes, and A. O. Effenberg, "A sound design for the purposes of movement optimisation in elite sport (using the example of rowing)," in *International Conference on Auditory Display*, Copenhagen, Denmark, June 2009.

[7] E. Summers, J. Langston, R. Allison, and J. Cowley, "Using SAS/GRAPH to create visualizations that also support tactile and auditory interaction," in *SAS Global Forum 2012*, Orlando, Florida, Apr. 2012.

[8] M. Friendly, "A brief history of data visualization," in *Handbook of Computational Statistics: Data Visualization*, Chen, C., Hrdle, Wolfgang, and Unwin, Antony, Eds. Heidelberg: Springer-Verlag., 2007, vol. 3, pp. 1–34.

A SONIFICATION PROPOSAL FOR SAFE TRAVELS OF BLIND PEOPLE

Pablo Revuelta Sanz

Carlos III University of Madrid,
Electronic Technology department,
Av. Gregorio Peces Barba 1, 28918, Spain
prevuelt@ing.uc3m.es

Belén Ruiz Mezcua

Carlos III University of Madrid,
Computer Science department,
Av. Gregorio Peces Barba 1, 28918, Spain
bruiz@inf.uc3m.es

José M. Sánchez Pena

Carlos III University of Madrid,
Electronic Technology department,
Av. Gregorio Peces Barba 1, 28918, Spain
jmpena@ing.uc3m.es

ABSTRACT

Sonification is one of the most natural ways to complete the information perceived by the blind people. Thus, it has been widely applied to create assistive products to help this community in their daily life. In our case, we are working on a mobility device which transforms the depth map of a scene into a set of sounds, comprehensible by the user. Our sonification proposal is based on the opinions of experts and potential users, collected by different interviews which crystallize in the herein explained sonification. This proposal follows the so-called *point transform*, which allows real-time sonification and quite accurate localization of the sound sources. However, some modifications to avoid ambiguous situations are also implemented and explained in this study.

1. INTRODUCTION

Sonification is the process in which some information is translated into sounds, formerly to ease the reception, but also for aesthetic or leisure purpose.

We are working on an assistive product called Assistive Product for an Autonomous Travel (APAT) [1], in which a sonification system helps blind people to mentally build a representation of his/her surroundings. For that purpose, we have developed an image processing step, in which two images are processed to obtain the depth map of the scene, by means of stereo vision techniques.

In this manuscript, we propose a novel and still-in-design process sonification code, which tries to surpass the limitations found in the bibliography, regarding this kind of technical aids for the blind community.

2. BACKGROUND

Sonification is a technique that has been widely used. The first prototype found sonifying images into sounds was built by Noiszewski, the Elektroftalm in 1897 [2]. Some years later, in 1912, d'Albe built the Exploring Optophone [3].

Since then, many assistive products have been proposed, especially in the last two decades. There are some basic dimensions into which any sonification proposal can be classified:

- Number of channels: Sonification using one channel (monaural emission) or two (stereo or binaural emission).
- Arbitrariness: Some sonification codes exploit the natural direction discrimination capability of the sounds. Others implement arbitrary codes for some parameters of the space, such as vertical position, which are not so well localized. There are some algorithms lying in the middle of these two groups.

We will focus on arbitrary and mixed options (no matter the number of channels, for instance), to summarize them into a few sonification paradigms. We will provide an example of a device implementing each paradigm.

- *Piano transform*: Height is codified as frequency, and horizontal axis as time. The brightness is correlated with the volume. That was d'Albe's choice.
- *Point transform*: Firstly proposed in [4], height is codified as frequency and horizontality as binaural loudness. The volume, again, is related to the brightness.
- *Pitch transform*: Proposed in [5], assigns the frequency to the depth (distance) of a point.
- *Verbal transform*: Extending he concept of arbitrariness, we can find projects as [6] where the surroundings are "read" by a synthetic voice.
- Other proposals: "Click" guiding of a Geiger counter with a radioactive emitter [7].

Regarding what should the light represent, we find 3 options:

- The visual brightness: Directly transforms the image into sounds, as it is done in [4]. This option is called *direct mapping*.
- The depth: The image is processed and only the depth is transformed into sounds.
- Edges: Only the edges are sonified, eliminating the volumes in the sonification process. An example of that transform is described in [8].

3. SONIFICATION

3.1. Information to be sonified

As said before, we have a depth map (usually called 2.5D image), gray scale, as shown in figure 1.

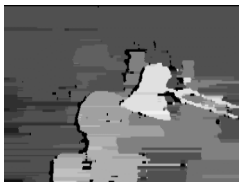


Fig. 1. 2.5D image example.

In this image, the brightness represents the distance of the point to the camera. The whiter a point is, the closer it is.

3.2. Sonification code: the Modified Point Transform

We decide to take as baseline the point transform (see section 2) over depth maps, because of the following reasons:

- The system must work in real time to avoid obstacles, so the time cannot be, directly, a variable of each image sonification.
- It uses the binaural properties of the human hearing system, making the training easier and more intuitive.

Important limitations were also found. For instance, this transform cannot differentiate between a volume centered in the image, and two bodies with half volume each, located laterally. Likewise, the volume may change with the ambient noise and, hence, it loses its capability of giving an absolute depth measure.

Thus, a final sonification code has been proposed, with the following characteristics:

- The brightness (the depth) is correlated with the volume, but the range of possible values is discretely split into 6 different sounds (synthetic voice, flute, oboe, trombone and muted trumpet), becoming sharper when the points become closer to help in distance discrimination.
- The lateralization is performed by differences in the loudness and the time of each sound, as it is described since the firsts psychoacoustic studies [9]. To avoid ambiguities, a tremolo is applied to lateral points, taking into account that the closer to a lateral a point is, the deeper is the tremolo.
- Only the nearest pixels (being their bright value higher than 42, in a range of [0,255]) are sonified.
- The vertical axis is codified by means of harmonic musical notes (which perform the CMaj7m chord when all the height levels are excited). However, some simpler profiles have also been proposed, being this last one the most complex. In this maximum level, 16 notes are used for height codification (the CMaj7m chord in 4 octaves). Any Harmonic chord allows the user to perceive music, instead of unpleasant noise.

The sonification is implemented by means of the MIDI standard protocol [10].

4. ACKNOWLEDGMENT

We acknowledge the student grant offered by the Universidad Carlos III of Madrid and CESyA.

5. REFERENCES

[1] P. Revuelta Sanz, B. Ruiz Mezcua, and J. M. Sánchez Pena, "ATAD: una Ayuda Técnica para la Autonomía en el Desplazamiento. Presentación del Proyecto" in *Libro de Actas: IV Congreso Internacional de Diseño, Redes de Investigación y Tecnología para todos*, pp. 151-160, 2011.

[2] W. Starkiewicz and T. Kuliszewski, "The 80-channel elektroftalm." *Proceedings of the lnternational Congress Technology Blindness, Am.Found.Blindness* . New York, 1963.

[3] F. d'Albe, *The moon element*, London: T. Fisher Unwin, Ltd., 1924.

[4] R. M. Fish, "Audio Display for Blind," *IEEE Tran. on Biomed. Eng.,* vol. 23, no. 2. pp. 144-154, 1976.

[5] E. Milios, B. Kapralos, A. Kopinska et al., "Sonification of range information for 3-D space perception." *IEEE Tran. on Neural Sys. and Rehab. Eng,.* vol. 11 no. 4, pp. 416-421. 2003.

[6] BESTPLUTON World Cie, "The "Mini-Radar", your small precious companion that warns you obstacles in a spoken way, and that helps you to walk straight." http://bestpluton.free.fr/EnglishMiniRadar.htm Apr. 2011.

[7] R. L. Beurle, *Summary of suggestions on sensory devices*, London: San Dunstan's, 1947.

[8] L. H. Riley, G. M. Weil, and A. Y. Cohen, *Evaluation of the Sonic Mobility Aid*. American Center for Research in Blindness and RehabiIitation, 1966.

[9] Lord Rayleigh, "On our perception of sound direction," *Philos. Mag.,* vol. 13. pp. 214-232, 1907.

[10] MIDI Manufacturers Association MMA, "General MIDI 1, 2 and Lite Specifications." http://www.midi.org/techspecs/gm.php 2012.

InteNtion – Interactive Network Sonification

Rudi Giot

LARAS - ISIB,
150, rue Royale,
1000 Brussels, Belgium
giot@isib.be

Yohan Courbe

LARAS - ISIB,
150, rue Royale,
1000 Brussels, Belgium
ycourbe@gmail.com

ABSTRACT

This paper presents an innovative approach in monitoring network traffic by adding a new dimension: the sound. InteNtion (Interactive Network Sonification) is a project aimed at mapping network activity to musical aesthetic. The network traffic analysis is made with the SharpPCap library (a port of WinPCap to C# environment). From this analysis, the collected data are converted into MIDI (Musical Instrument Digital Interface) messages and sent to dedicated synthesizers, which generate sounds dynamically mixed together. The whole process results in an interactive soundscape. This novel approach will initiate two opportunities for technological development. It allows users to actively take part in an interactive exhibition system through simple actions involving network access, including streaming radio over the Internet, sharing music on Twitter, downloading mp3 files and others. This project initiates also a new dimension in monitoring the network by helping the administrator in detecting efficiently the hacking and abuse of the infrastructure.

1. INTRODUCTION

Networked systems are more and more ubiquitous at home, at work, in bars, shops … even in streets. We are facing a growing number of hotspots and services, allowing us to stay connected any time of the day.

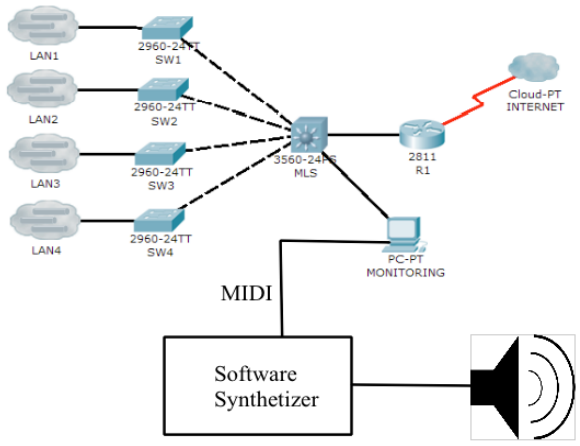


Figure 1: Traffic analysis and sonification

The problem is that, because of the number of users (who are anonymous most of the time), the networks remain more and more vulnerable to hacking or usage abuse. To avoid that, professionals use a lot of different softwares to monitor all the aspects of their network, providing some text alerts, statistical graphics and so on. Thus, the traffic control is essentially based on visual reports.

We propose to add to this existing dimension a new one which is based on sound. The initial idea of this project is to analyze the network flow via a software program, called NAC [8] (Network Access Control) and map these data into a network sonification.

Several projects propose to sonify Internet sites [3] or to focus on specific parameters such as the network performance [4]. Our approach is different. The solution we propose is to divert the standard applications' libraries to analyze the traffic at a very low level (layer 3 of the ISO model [2]) which brings us to a new way of interacting with the network. Based on other studies [5] and [6], we have imagined mapping the data into MIDI events, which are sent to software music synthesizers. The output is a soundscape literally composed by the network activity. It means that users can act interactively with the sounds just by surfing on the Internet, sharing data on Facebook or synchronizing their Dropbox. You will definitely hear your network in another way.

2. THE CONCEPT

Applications such as WireShark [1] are useful to monitor a network, to detect intrusion, hacking or bugs. Those programs are mainly based on the PCap library that can be integrated in other projects for different, original and artistic purposes. We propose a software solution that analyzes live streams of an entire network flow (this software is running on the PC-PT-Monitoring; Figure 1) and generates usage statistics of the network. The Internet Protocol (IP) parameters (including TCP/UDP segment and the data) are manifold: the Time-To-Live (TTL), the packet size, the fragmentation information, the source and destination addresses, the type of service and all the port usage statistics included in the TCP/UDP segments.

These collected data are then converted into MIDI messages to be sent on a MIDI interface as “Note On” or “Control Change” events on different channels (Figure 2).

At this stage of our work, the mapping is still experimental. However, using the method described in the following paragraph, the first musical aesthetic results are very promising.

We use four different sound synthesizer softwares from Native Instruments [9]. Taking the protocol into account, an analysis of the datagram is effected. We send the MIDI message to the associated synthesizer depending on the protocol (HTTP, FTP, DNS and for the fourth synthesizer all the other existing protocols). The amount of useless datagrams (e.g., ACK packet) considered as "noise" for the communication will be mapped in relation to the amount of "noise" in another particular granular synthesizer.

The other parameters are mapped as follows :

- The packet size is mapped to the frequency. A small packet gives a high frequency and the large one will result in a bass sound.
- The TTL of the datagram is used as duration of the note.
- The bandpass of the network, measured in real time, will modify by a Midi Control Change the bandpass of a resonant filter in the appropriate synthesizer.
- We can estimate by the IP addresses, the distance between the origin and the destination. That parameter will change the reverb size of the sound with a Midi Control Change.

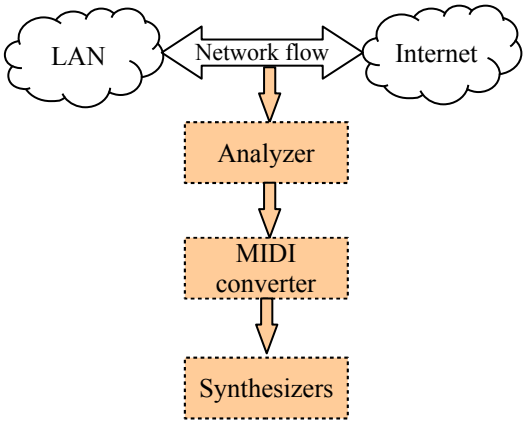


Figure 2: The different layers of the project

Thus, the synthesizers produce sounds mixed together which result in a soundscape. The soundscape will react on what is going on the network [7]. People using the network will, due to their own actions, modify the activity and consequently the soundscape.

3. CONCLUSION

At this stage of our work, InteNtion provides an unusual and innovative way to monitor networks by using their entire data flow to create musical aesthetic. Therefore we can consider this project as an artistic project that might be used, for instance, in a concert where the audience participates as main composer in the music performance.

A second result of this project is the initiation of a software program allowing a network administrator to monitor the infrastructure in differentiating regular, normal data flow from an unusual behavior (abuse or hacking) just by hearing sounds.

4. REFERENCES

[1] http://www.wireshark.org/, 2012
[2] Andrew S. Tanenbaum and David J. Wetherall, *Computer Networks*, Prentice Hall, 2010
[3] Zach Layton, *Network Sonification*, http://turbulence.org/Works/net_sonification/, 2007
[4] Chris Chafe and Randal Leistikow, *Levels of temporal resolution in sonification of network performance*, Proceedings ICAD, 2001
[5] Diniz, Deweppe, Demey, Leman, *A framework for music-based interactive sonification*, Ghent University, 2010
[6] Thomas Hermann, *Sonification for Exploratory Data Analysis*, Universität Bielefeld, February 2002
[7] http://www.youtube.com/watch?v=xWq24O0lU_g, 2012
[8] http://nac.dev.isib.be/, 2011
[9] http://www.native-instruments.com/, 2012

INTERFACING THE EARTH

Peter Beyls

University College Ghent,
School of Arts,
Jozef Kluyskensstraat 2, B 9000 Gent, Belgium
peter.beyls@hogent.be

ABSTRACT

We provide a short introduction to *WindChime*, a real-time web-driven audiovisual installation. Weather data from many world locations is gathered from a server and accommodated in a dynamic visual representation. The dynamics of the wind at specific world locations exercises influence over a mass of floating particles in a virtual parallel world. Particles in turn influence the production of complex sounds. In effect, a rewarding aesthetic experience results from the appreciation of the intricate interplay of two complex dynamical systems; one of natural origin (the earth), the other of cultural design (the program).

1. INTRODUCTION

Artists developing private first principles might suggest new scales in time and space while challenging the notion of dimensionality, both conceptually speaking and in terms of embodiment. This includes the exploration of sound aiming the expression of spatiotemporal complexity hidden in a tiny organic micro-world [1]. In contrast, project *WindChime* suggests viewing the whole Earth as a dynamic system subject to sonification [2]. In essence, we implement a virtual version of the archetypal wind chime; an arrangement of objects suspended from a frame creating tinkling sounds in a light breeze.

Previous research exploring the Earth as a global source of information includes the translation of the *Kp* indices reflecting the Earth's magnetic field into musical pitches and compressing thousands of data items into a few minutes of musical time [3]. *Sonification / Listening Up* is a more recent MIT project aiming the sonification of the interplay of sun winds with the Earth's atmosphere, a continuous interaction that takes place some 60 miles above ground level [4].

The conviction that rewarding aesthetic experiences may result from the perception of multifaceted behavior in a given complex system underpins the present project.

More precisely, the global systems output here emerges from the confrontation of two complex dynamical systems: (1) the complex stretch of non-linear forces instructing the development of wind across the surface of the Earth, and (2) the largely unpredictable (though coherent) behavior in a sounding network of digital audio processing units. So, the earth is considered a *found system* while the sound producing system is a deliberately constructed system; the net result is collaborative effort involving a natural and a cultural system.

2. IMPLEMENTATION

The project is conceived as a real-time web-driven audiovisual installation. The implementation continuously captures the intensity and direction of the wind at many different locations worldwide by probing live data from a server at the National Center for Atmospheric Research [5]. Implementation consists of two concurrent programs, (1) a Java program running the web sensing functions, the dynamic visualization and the analysis and mapping functions and (2) a program written in SuperCollider [6] handling real-time audio synthesis. The programs communicate through OSC [7].

The Java program holds a number of classes from which functional objects are instantiated: the *World* includes *Particles*, their behavior being influenced by forces emanating from a *Field*, the strength of the Field being developed on a continuous basis from local data gathered from *Stations* providing live weather information. A brief description of the functionality inside every class follows.

The Stations class holds a data structure containing information on 7961 weather stations. A single 80-character entry contains 18 data items, including name of location, a four-character international ID, latitude and longitude, elevation, aviation specific information and country code, for example:
ISLE OF MAN/RONA EGNS 03204 54 04N 004 37W 17 X T 6.

A single *Station* object is instantiated by randomly selecting a candidate station from the list of potential stations. The Station object computes its visualization on a world map image - displayed as a permanent background image – by converting its latitude/longitude data to a Cartesian map (see figure 1). In addition, the object makes a request to the server and, when available, parses the data received for extraction of wind strength and wind direction at that station's location.

A single global *Field* object holds two complementary matrixes (32 by 20 elements) called *data* and *previousData* – they hold information about the strength of the wind captured for the whole world over a span of two consecutive time frames. The representation of the matrix is actually mapped on top of the world map – respective matrix locations are imagined as being connected to specific physical locations in the world map.

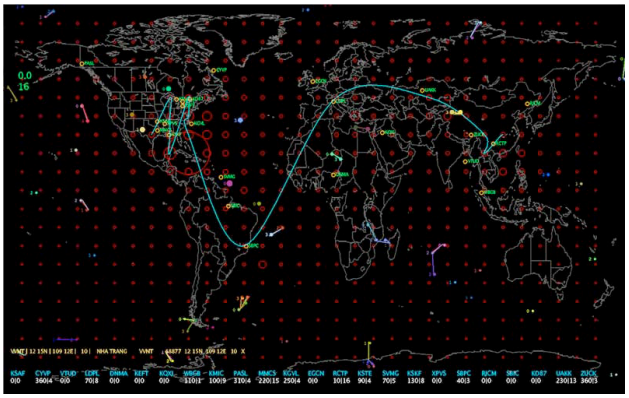


Figure 1: Snapshot of visualization resulting from sampling 24 world locations for real-time weather data.

The data gathered from the current Station in the World updates the Field at a specific location i.e. where the (normalized) image of the matrix and the station's location coincide. In addition, the numeric data in every single matrix element is scaled down in proportion to its distance from the current Station XY location. In the long run, following successive process cycles, the matrix will dynamically capture the strength of the wind with live data from 24 stations simultaneously. The World object actually aims to optimize data input and gradually builds a data structure of locations in the world trying to maximize the effect of the wind in the installation as a whole. In the current implementation, the matrix is visualized as an array of red circles, their radius relative to the strength of the wind at the respective locations.

A *Particle* is envisioned as virtual dust facing – a representation of – actual physical wind. Particles float in 2D space, their velocity and size is modulated by the intensity of the Field being expressed at their respective positions. Particles are also sensitive to their surroundings: neighboring particles within a given critical distance produce temporary clusters visualized by line segments. Particles within clusters interact in two ways, (1) a particle will adapt its angle of movement to the angle of one of its (randomly selected) neighbors and (2) a particle's energy level will boost in proportion to its number of neighbors. An isolated particle (no neighbors) will slightly decrement its energy level in every process cycle, energy levels are considered in the audio mapping algorithm documented in section 3.

A single *World* object accommodates 100 particles. The World creates a list of 24 unique stations that return actual data (not all servers are operational on a permanent basis). The data from all 24 stations is visualized and accommodated in the field matrix. Every station remains active for some time interval (normalized to a scale; from 30 seconds to 5 minutes) in proportion the strength of the wind at its specific location.

Figure 1 shows the world map in the background, the Field matrix (the red circles reflecting the local intensity of the wind) and a few clusters of floating particles. The blue curve is computed by interpolating between data points above a given threshold, the curve thus an emergent phenomenon built by forces spread out around the globe.

3. MAPPING

The mapping strategy developed here is unusual as it aims to develop a sensible association between behaviors in two independent parallel systems that coexist within their private domain. This procedure attempts to avoid the simplistic notion of conventional mapping [8] or direct sonification [9]. Audio synthesis in *WindChime* explores the principle of "influence" as detailed next.

A complex audio network is developed – by trial-and-error method, much like trying patches on an analog synthesizer – by patching a critical collection of synthesis and processing modules. Audio complexity builds up because the modules interact in non-linear ways and, given certain parameter settings, the global synthesis engine engages in chaotic behavior. Although the patch remains static, it reveals a quite significant expressive musical space. In addition, the patch can be pushed into a great many behavioral modes, its operational integrity remains guaranteed and its sonorous identity equally recognizable. The patch is characterized by control economy: it has only two entry points for external signals, so it may be imagined as a 2D surface accepting a single XY location. Inside a patch, X and Y control signals map to many different parameters simultaneously, however using distinct interpreter algorithms. The mapping strategy is consequently minimal on the side of "control" (only 2 parameters), yet the system aims to maximize its effect on audio complexity through the critical design of a networked synthesizer.

Now, the free running audio patch continuously consults the Field instance variable of the World. Any particle may trigger a sound when its present location (i.e. the contents of the Field at the particle's location) exceeds a given adaptive threshold. The threshold increases while facing overstimulation; the absence of input (e.g. "wind energy") will lower the threshold thus increasing the probability of audio responses. The adaptive algorithm actually contributes to global emergent behavior in *WindChime*. In addition, this project features real-time visualization of pinged information from the weather data server, the station's ID's are displayed and the accumulated Field is stretched across the world map. Interaction between particles shows up in dynamic computer animation.

4. CONCLUSION

The present project bridges two complex dynamical systems: the progression of wind patterns around the globe with the development of audio patterns inside a complex digital audio patch. Aesthetic appeal follows from the perception and appreciation of the complementary complexity inside the unfolding visual representation of the world's wind data in relation to the unfolding sonorous complexity enacted by the audio synthesis patch. The *WindChime* project suggests evidence that fractional recognition of relationships between behaviors in both systems provides the basis for a rewarding human-machine experience.

5. REFERENCES

[1] E. Miranda, A. Admatzky and J. Jones, "Sound Synthesis with Slime Mould of Physarum Polycephalum". Journal of Bionic Engineering, 8: 107-113, 2011

[2] G. Kramer (ed.) Auditory Display, Sonification, Audification, and Auditory Interfaces, Addison-Wesley, Reading, MA 1994

[3] C. Dodge, The Earth's Magnetic Field, Nonesuch Records, 2-track LP, H71250, 1970

[4] L. Heineman http://www.mit.edu/spotlight/ionosphere/ 2005

[5] UCAR 2012: http://www.rap.ucar.edu.

[6] S. Wilson, D. Cottle and N. Collins, *The SuperCollider Book*, The MIT Press, Cambridge, MA 2011

[7] M. Wright, "Brief Overview of OSC and its Application Areas", OSC Conference, Berkeley, CA 2004

[8] J. Chadabe, "The limitations of mapping as a structural descriptive in electronic instruments", NIME '02 Proceedings, Singapore, 2002

[9] A. de Campo, J. Rohruhuber, T. Boverman, and C Frauenberger, "Sonification and Auditory Display in SuperCollider". In: The SuperCollider Book, The MIT Press, Cambridge, MA 2011

IMPOVING THE EFFICACY OF AUDITORY ALARMS IN MEDICAL DEVICES BY EXPLORING THE EFFECT OF AMPLITUDE ENVELOPE ON LEARNING AND RETENTION

Jessica Gillard

McMaster University,
McMaster Institute for Music and the Mind,
1280 Main St. W. Hamilton, ON, Canada
gillarj@mcmaster.ca

Michael Schutz

McMaster University,
McMaster Institute for Music and the Mind,
1280 Main St. W. Hamilton, ON, Canada
schutz@mcmaster.ca

ABSTRACT

Despite strong interest in designing auditory alarms in medical devices, learning and retention of these alarms remains problematic. Based on our previous work exploring learning and retention of associations between sounds and objects, we suspect that some of the problems might in fact stem from the types of sounds used. Several of our previous studies demonstrate improvements in memory associations when using sounds with "percussive" (i.e. decaying) envelopes vs. those with "flat" (i.e. artificial sounding) envelopes – the standard structure generally used in many current alarms. Here, we attempt to extend our previous findings on the effects of temporal structure on the learning and memory. Unfortunately, we did not find evidence of any such benefit in the current study. However, several interesting patterns are emerging with respect to "confusions" – the times when one alarm was confused with another. We believe this paradigm and way of thinking about alarms (i.e. attention to temporal structure) could provide insight on ways to improve auditory alarms, thereby prevent injuries and saving lives in hospitals. We welcome the chance to gather feedback on our approaches and thoughts as to why our current attempts (which we believe are based on a solid theoretical basis) have not yet led to our hoped-for improvements.

1. INTRODUCTION

Medical alarms are designed to alert medical staff immediately when there is a problem with a patient. Despite their ability to grab attention, the current design of medical alarms is ineffective and results in several deaths and injuries among patients each year [1]. This could be attributed, in part, to the poor learning and high rates of confusion seen in empirical tests of alarm learning [2,3].

Confusions among alarms can arise from a variety of factors such as acoustic similarity, functional similarity or difficulty forming associations [1,4]. The current study aims to investigate a more subtle change to the current acoustical design of medical alarms. Previous studies in the lab have focused on the ecological validity of sound and its effect on associative memory abilities. The main manipulation in these studies was the shape of a sound over time; or more technically known as the 'amplitude envelope'. Two types of amplitude envelopes were investigated: flat and percussive (as shown in Figure 1A). A percussive envelope is representative of impact sounds, which we hear on a regular basis. A flat envelope, on the other hand, is man-made and is heard significantly less often than percussive sounds. Therefore we hypothesized that percussive sounds could be learned and recalled much easier than flat sounds due to our extensive experience with them. This hypothesis was supported in several object-melody association experiments [5], where it took significantly fewer trials to learn the associations when percussive melodies were used compared to flat melodies. Here, we will investigate the role of amplitude envelope as it applies to the learning and memory of medical alarms.

2. METHOD

2.1 Participants

Participants consisted of 48 undergraduate students (16 Male, 31 Female, 1 Transgendered) ranging in age from 17 to 26 (M = 19.06, SD = 1.80) recruited from the undergraduate psychology and linguistics pools

2.2 Stimuli and Apparatus

We selected eight tone sequences from a set used in a previous study [5]. In order to create both flat and percussive versions of the tone sequences, SuperCollider[1] was used to shape pure tones (i.e. sine waves) into flat and percussive envelopes to create individual tones. These individual tones were then arranged into sequences using Audacity[2] - a free sound editing program. All tone sequences consisted of four one-second sound clips that were either all percussive or all flat concatenated together to make a four-second track. Percussive tones were approximately 800ms in length and were separated by 150ms. Flat tones were 745ms in length and were separated by 200ms. Each of the tone sequences were labeled with a number from 1 to 8 and are shown in Figure 1B.

Tone sequences were stored on an iMac computer and presented over Sennheiser HDA 200 headphones at a comfortable listening level, which was held constant.

A) B)

[1] http://supercollider.sourceforge.net
[2] http://audacity.sourceforge.net

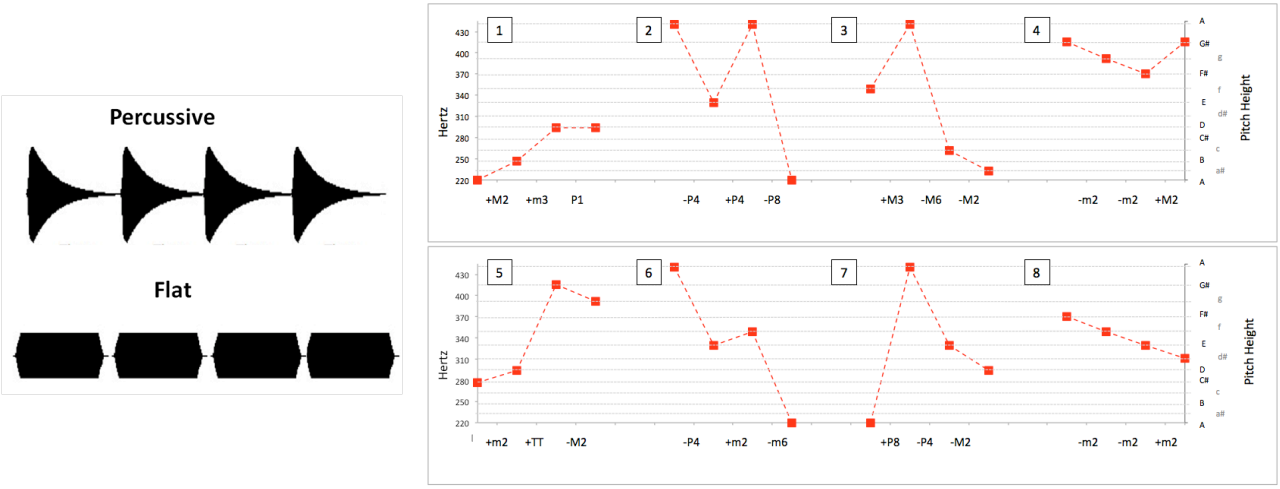


Figure 1 – Percussive and flat waveforms (A) and the contours of the 8 melodies used in the Experiment (B).
M = Major, m = Minor, P = Perfect, TT = Tritone, + = Ascending, - = Descending.

Participants filled out a short survey that included questions regarding demographics, musical training and musical practice and listening behaviours prior to beginning the experiment.

2.3 Procedure

To control for differences I musical training, participants heard alarm sets that contained four flat and four percussive tone sequences. The pairings tone sequences and alarm labels were randomized for each participant.

The experiment consisted of four phases: Study, Training, Break and Evaluation. During the Study phase, participants heard each of the 8 alarms twice and were told its correct alarm association. In the Training Phase, we presented participants with each of the tone sequences once (in a random order) and asked participants to identify the correct alarm association. Participants were given feedback on their correctness, played back each of the sequences and were told the correct alarm association regardless of their answer. Participants heard all 8 melodies, which made up a block of training. These blocks were repeated until participants could correctly identify 7 out of 8 tone sequences in 2 consecutive blocks, or reached a maximum of 10 blocks. Once the Training phase was complete, subjects took a five-minute Break to play a mini-golf game on a computer. The sound was turned off to ensure the game sound effects did not interfere with our evaluation of learning and retention of the tone sequences. During the Evaluation phase, participants were randomly presented with each of the 8 tone sequences and were asked to identify the correct alarm association. Participants received their final score upon completion. This paradigm is a hybrid pairing of one used in previous studies [5], and what is currently used in medical alarm research [2,3].

3. RESULTS AND DISCUSSION

Performance on flat and percussive alarm associations in the Evaluation phase was compared using a pair-wise sample T-test and yielded no significant difference ($p = 0.554$).

We are currently examining the patterns of confusion and suspect amplitude envelope, timbre and tonality play significant roles. To our knowledge, no studies have looked at alarm confusions based on these aspects and think this might provide some insight on ways to improve auditory alarms.

Given the importance of alarms in a medical setting, the design of sounds that can easily be associated with their intended meaning is a pressing and timely issue. Therefore, we are very interested in any comments and feedback on the research presented in this extended abstract.

4. ACKNOWLEDGMENT

This work was supported in part through grants from the Natural Sciences and Engineering Research Council - RGPIN/386603-2010, and the Early Researcher Award - ER10-07-195 (Michael Schutz, PI).

5. REFERENCES

[1] J. Edworthy and E. Hellier, "Alarms and human behaviour: implications for medical alarms," *British Journal of Anaesthesia*, vol. 97, no. 1, pp 12-17, July 2006.

[2] P.M. Sanderson, A.N. Wee, and P. Lacherez, "Learnabiliy and discriminability of melodic medical equipment alarms," *Anaesthesia*, vol. 61, no. 2, pp 142-147, February 2006.

[3] A.N. Wee and P.M. Sanderson, "Are melodic medical alarms easily learned?," *Anesthesia & Analgesia,* vol. 106, no. 2, pp 501-508, February 2008

[4] J. Edworthy, E. Hellier, K. Titchener, A. Naweed and R. Roels "Heterogeneity in auditory alarm sets makes them easier to learn," *International Journal of Industrial Ergonomics,* vol. 41, no. 2, pp 136-146, March 2011.

[5] M. Schutz, J.K.Stefanucci, A. Carberry and A. Roth, "Name that (percussive) tune: Tone envelope affects learning," *in Proc. of the (nth) Int. Conf. of Music Perception and Cognition*, Long Beach, USA, 2007.

EFFECTS OF PLEASANT AND UNPLEASANT AUDITORY MOOD INDUCTION ON THE PERFORMANCE AND IN BRAIN ACTIVITY IN COGNITIVE TASKS

Matti Gröhn, Lauri Ahonen, Minna Huotilainen

Finnish Institute of Occupational Health
Topeliuksenkatu 41 a A, FIN-00250 Helsinki, Finland
firstname.lastname@ttl.fi

ABSTRACT

Our study focuses on mood induction with pleasant and unpleasant auditory stimuli during the break. Our test includes subjective evaluation (NASA-TLX, KSS, POMS), cognitive tests and brain responses (MEG and EEG). We aim studying the effect affective state has on work-like tasks. Hypothesis: pleasantness of auditory mood induction affects cognitive performance and brain responses.

1. INTRODUCTION

Cognitive performance results from multiple factors including arousal and physical fatigue. Affective state, major modulator for cognitive and physical performance, is often neglected. Here we investigate the effects of different dimensions in affective state on performance in cognitive tests and electrophysiology.

Music has the indeniable ability to convey strong emotions. In addition, lots of research under the topic of music is devoted to revealing the possible effects music claimed to have on cognitive performance. One of the famous cases is the controversial Mozart effect[1]. Not more than decade ago there was an intense debate about the reputed effect Mozart's music has on performance in certain spatially demanding cognitive tasks. The debate more or less concluded that music has an effect on affective state. This effect is called enjoyment arousal, i.e., arousing and positively valenced effect of musical experience[2]. However, the research around the discussion did not try to specify the dimensions of affective state influenced by the effect, in terms of the two dimensional model namely valence vs. arousal. A novel paradigm is designed for measuring the effect of mood on performance and related electrophysiology in cognitive tests. The tests are chosen to simulate a cognitively demanding tasks during a regular workday in office like work environment.

2. MOOD INDUCTION

There are a number of experimental techniques that have been developed to induce positive or negative mood in the participants. The effectiveness of these Mood induction procedures (MIP) has been investigated [3] and the MIPs using music have been shown to be effective. For example, the cardiovascular and respiratory patterns are changed according to the mood induced by music [4]. When studying the physiological features related to changes of mood, e.g., measures of heart rate or heart rate variability, blood pressure, etc., the acoustic and especially rhythmic content of the musical material plays a key role. One may expect effect related to temporal synchronization of the physiological functions and the musical material. Such synchronization, or entrainment of especially respiration to the temporal characteristics of the music used in the mood induction may completely override all physiological effects in the study [5].

Questionaires/Preparations	Personal Info + POMS	20 min
Testset 1	Cognitive tests	35 min
Questionaires	NASA-TLX, KSS, POMS	5 min
Mood induction	Pleasant/Unpleasant	12 min
Questionaires	POMS	5 min
Testset 2	Cognitive tests	25 min
Questionaires	NASA-TLX, KSS, POMS	5 min

Table 1: One visit testprotocol.

3. METHOD

3.1. Subjects and Procedure

We have currently measured 7 subjects. Five of them were female and two male. Their ages varied from 21 to 38, mean 27 standard deviation 6.6. In Table 1 is a protocol for one visit. Each visit starts with questionaires and preparations for the MEG and EEG measurements. The Profile of Mood States (POMS) are used to measure affective mood state of the subject.

We have three cognitive tests in our protocol: N-back, Task Switching and Image Memory. In Testset 1, each test includes training trials before the measurements. In Testset 2, tests are accomplished without training. The testsets are performed before and after a mood induction in two different days, i.e. tests are performed four times in total. Paradigm is counter balanced so that half of the subjects get negative induction first as the other half get positive treatment first.

After the cognitive tests the NASA-TLX, KSS and POMS questionaires are accomplished.

After the first round there is a break, which includes pleasant or unpleasant mood induction. In pleasant mood induction subjects are exposed to music of personal choice. We provide preselected playlists from different genres among pop, electro, classical, and rock. The playlists were created considering the effect on arousal. Mean tempo of each playlist is adjusted to match the mean tempo of the unpleasant sound sequence. It is mix of environmental noise created by superimposing negatively assosiable sounds, e.g. crying, alarms, and dental drills.

In the MEG we obtained a large number of brain responses from which we determined their amplitude and timing, especially

	KSS	Mental Demand	Frustration
Before Pleasant	6.2 (1.5)	2.0 (1.4)	3.0 (1.0)
Before Unpleasant	5.3 (1.7)	2.3 (1.6)	2.7 (1.8)
After Pleasant	5.7 (1.9)	2.5 (1.1)	2.5 (1.9)
After Unpleasant	5.5 (0.8)	2.3 (1.6)	2.3 (1.2)

Table 2: Mean and standard deviation for KSS (scale 1 (aroused) to 7 (sleepy)) and NASA-TLX (scale 1 to 5).

the onset and peak times. The responses are elicited by the presentations of the figures and sounds that occur in the cognitive experiments. In all test types the responses, reaction times, stimulus onsets and related electrophysiology are recorded with Neuromag data acquisition system.

N-back task: The participants will complete a visual n-back task after a practice trial. In each condition a total of 180 stimuli (numbers) will be presented one at a time on a computer screen. The participants will be instructed to give a response to whether the number is the predetermined "x" within a sequence of numbers (0-back), the same as the previous number (1-back), or the same as the number presented 2 numbers back (2-back), or not.

Task switching. When people have to switch between two tasks, they are slower on the task-switching than on the task-repeating trials [6]. We compared the time needed and mistakes made in unpredictable task-switching trials to those in task-repetition trials.

Image Memory. Tests for visual memory are prone to errors hence a good assessing technique for cognitive performance [7]. The test is alternation of The Cambridge Face Memory Test [8]. Instead of faces we use abstract images selected in collaboration with clinical psychologists. The difficulty of the task is adjusted by adding images to target train to achieve sufficient error frequency.

4. RESULTS

The questionnaire about the pleasantness of the break and its effects showed that the subjects found the break with self-chosen music to be clearly more pleasant than the break with noise. The perceived pleasantness of the break in scale 1 to 5 was 3.6 (0.24) for pleasant and 1.3 (1.2) for unpleasant break. The KSS questionnaire showed that the break changed the arousal level: the pleasant break increased it and the unpleasant break decreased it. In the NASA-TLX, the mental demands and the frustration changed differently according to break type, see Table 2. In POMS, the effect of the break type was most clearly seen in the arousal dimension. In the MEG and EEG, all three test types produced clear responses. For example, the presentation of the task image in the task switching experiment gave rise to clear P300 responses. The response magnitudes and latencies in the four conditions were compared. Simultaneously, the performance data from the experiments were obtained. The intra-subject variability was large compared to the effects of the break type.

5. DISCUSSION

Inducing mood changes by listening to music or sounds has been demonstrated in many experimental set-ups. Our aim was to induce mood changes by music or sounds during a short break between cognitively demanding tasks. We succeeded in creating a pleasant and an unpleasant break with measurable effects. When changes in the mood states, sleepiness and task-related effects were compared after the pleasant and unpleasant break, the results were individually quite variable and dependent on the individual situation prior to the break. The task performance of the participants became better during the experiment - the reaction times got faster and the hit rates became generally better. Such a learning effect has been shown also in previous studies. The effect of the short break in the midst of the task performance was surprisingly small and was largely masked by the learning effect. This may be due to the fact that the tasks themselves also induced changes in the mood. The tasks are demanding and frustating and may thus override the effects of the pleasant break. The EEG and MEG methods are clearly capable of capturing the brain activity related to the three tasks since clear brain responses were observed. The participants' response patterns were affected by the breaks, but the effects were not clearly consistent across subjects. This may be due to the differential effects of the tasks on individual subjects' mood. In sum, mood induction with music or sounds was found to be possible even during short breaks and it may affect the cognitive task performance.

6. ACKNOWLEDGMENT

We thank our subjects and BioMag personnel. In addition, we thank the reviewers for their valuable comments, which improved our paper.

7. REFERENCES

[1] F. H. Rauscher, G. L. Shaw, and K. N. Ky, "Music and spatial task performance," *Nature*, vol. 365, p. 611, Oct 1993.

[2] E. G. Schellenberg and S. Hallam, "Music listening and cognitive abilities in 10- and 11-year-olds: the Blur effect," *Ann. N. Y. Acad. Sci.*, vol. 1060, pp. 202–209, Dec 2005.

[3] R. Westermann, K. Spies, G. Stahl, and F. Hesse, "Relative effectiveness and validity of mood induction procedures: a meta-analysis," *European Journal of Social Psychology.*, vol. 34, pp. 557–580, 1992.

[4] J. Etzel, E. Johnsen, J. Dickerson, D. Tranel, and R. Adolphs, "Cardiovascular and respiratory responses during musical mood induction," *International Journal of Psychophysiology*, vol. 61, pp. 57–69, 2006.

[5] C. Wientjes, "Respiration in psychophysiology: methods and applications," *Biological Psychology.*, vol. 34, no. 2–3, pp. 179–203, 1992.

[6] R. Rogers and S. Monsell, "The costs of a predictable switch between simple cognitive tasks," *J. Exp. Psychol. Gen.*, vol. 124, pp. 207–231, 1995.

[7] B. J. Sahakian, R. G. Morris, J. L. Evenden, A. Heald, R. Levy, M. Philpot, and T. W. Robbins, "A comparative study of visuospatial memory and learning in Alzheimer-type dementia and Parkinson's disease," *Brain*, vol. 111 (Pt 3), pp. 695–718, Jun 1988.

[8] B. Duchaine and K. Nakayama, "The Cambridge Face Memory Test: results for neurologically intact individuals and an investigation of its validity using inverted face stimuli and prosopagnosic participants," *Neuropsychologia*, vol. 44, pp. 576–585, 2006.

SONIFICATION FOR THE ART INSTALLATION DRAWN TOGETHER

Mason Bretan, Gil Weinberg, and Jason Freeman

Georgia Institute of Technology
Atlanta, Ga
masonbretan@gmail.com gilw@gatech.edu jason.freeman@gatech.edu

ABSTRACT

This extended abstract describes Drawn Together, an interactive art installation in which a person takes turns drawing with a computer. We describe the process of the interaction and the methods used to creatively sonify the process and the animations. There are three main states in the interactive process that are sonically represented using audio samples in a mix of background and foreground sounds. The lines drawn by the computer are sonified using a set of features describing length, rate of time drawn, location, and curviness.

1. THE DRAWN TOGETHER PROJECT

Drawn Together is an installation art piece in which an individual and computer draw in a turn taking interaction. It was developed by the Open Ended Group in collaboration with the Georgia Tech College of Architecture and the Center for Music Technology. A camera, projector, two computers, four microphones, and numerous LEDs are encased in a table designed specifically for the piece. An individual is encouraged to draw on a single black sheet of paper. The computer responds with a 3D projection on to the same paper based on an analysis of the person's drawing. The participant responds to the computer's drawing with additional drawings on the paper, the computer responds again, and the process continues as a conversation unfolds between participant and computer via the shared drawing surface. There are three primary states to the piece: 1) the human drawing state 2) the "thinking" state and 3) the computer response state. The entire event includes an auditory component that enhances the experience through a sonification of the drawing and the state of the system (in terms of the three states). The audio is played through a pair of headphones worn by the individual currently interacting with the installation. There are also two loudspeakers on either side of the table allowing everybody else in the room to experience the sound.

2. TABLE DESIGN

The design of the table was influenced by the notion that in addition to a drawing surface the table could be a musical instrument. We consulted with a luthier to determine how best to achieve this and created a structure that acoustically manipulates the sound of the drawing implement on the table. The elongated table top serves as a resonating body and is filled with honeycomb shaped boxes, which filter the sound. Each honeycomb is an enclosed box with an F-hole at the bottom. The F-holes differ in both size and location for each honeycomb allowing for subtle auditory variation. Additionally, different size boxes add a variety of filtering characteristics. The sound becomes dependent on the implement being used (chalk, pen, pencil, pastel, etc) and on its point of contact with the table.

This work was supported by NSF Grant #0905516, Georgia Tech College of Architecture, School of Architecture, School of Industrial Design, School of Interactive Computing, and the Center for Music Technology

Figure 1: An individual interacting with *Drawn Together* as others observe

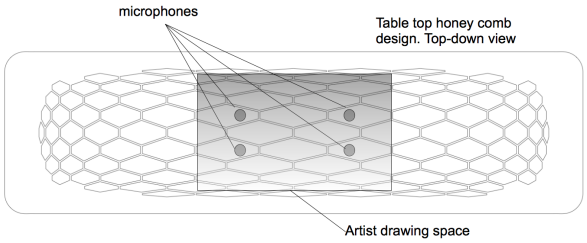


Figure 2: Honey comb design of the table top and microphone placement

3. SONIFCATION

The purpose of the audio is to provide a sonic component that is beautiful, creative, and relevant to the visuals, the state of the system, and the process of drawing. To achieve this we use a combination of techniques including a persistent background drone throughout and a foreground layer having different textures specific to each state.

3.1. Background

The background sound sets the tone for the entire installation. It was created by taking a recorded sample of a person drawing on the table, convolving the sample with an impulse response of the table, and convolving that result with a bowed cello note. The result of this process is a low drone which is present throughout. The drone changes pitch when convolved with a new cello note. A change in pitch signals a state change to a different phase of the interaction. This change in pitch sonically shifts the sound and also gives outside listeners not currently interacting with the table a sense of the rhythm of the interaction between the computer and the individual.

3.2. Foreground

In contrast to the background, the foreground audio was developed to give a more precise representation of the drawing activity of both the participant and the computer. This needed to be accomplished while still being aesthetically pleasing and maintaining the gentle ambiance desired.

In State 1 the person drawing hears the sound of his or her drawing implement on the paper. The sound is amplified and slightly filtered to soften the undesirable frequencies. The four microphones embedded in the table are located directly below the paper and spaced so that the output creates an auditory spatialization identical to the implement's location relative to the center of the paper. The two microphones on the right are mixed and sent to the stereo right channel and left side microphones are sent to the left channel. In our original implementation we used binaural filters to create a 3-dimensional sound spatialization. This was an attempt to further make the listener feel as if he were sonically immersed in the experience with the sound revolving around his head along a horizontal plane. Though after listening, we concluded that the binaural filters reduced the quality and effectiveness of the natural sound of the implement on the table.

State 2, the "thinking" state, is the interval of time between the point the person finishes drawing and the point when the computer starts its response. During this time the computer is analyzing the drawing and determining an appropriate response. In addition to the background pitch change a sample of a slow ticking clock is played to indicate the state.

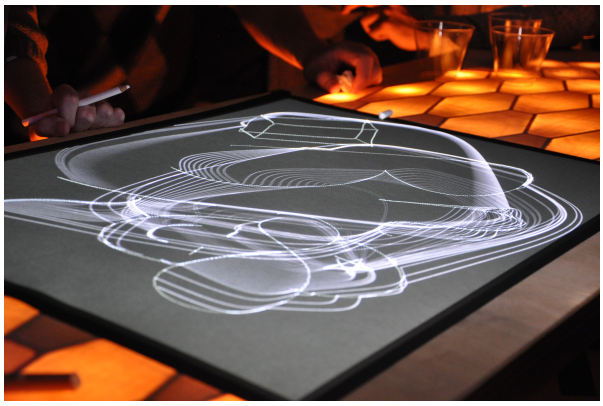


Figure 3: An example of a computer generated response with a person's drawing

State 3 provides the most challenging and interesting part of the sonification process. The foreground audio sonifies the computer's response in real-time. The current iteration of the music uses a large library of sounds taken from recordings of several human drawing gestures which were classified into groups such as straight lines, curvy lines, dashed, dotted, etc. In addition to these drawing sounds we also recorded and labeled sounds from a rain stick, wooden drum, and guitar. These samples are mapped to the 3-dimensional lines the computer draws. A large set of features describing each line and the density of lines being drawn at one time determines which samples to use. Some of the individual line features include location, length, and rate at which it is drawn. Each line is defined as a Bezier curve so it is also possible to get a measure of "curviness." Depending on the sample the algorithm selects, the sample is either looped or the playback speed is adjusted to be the same duration as it takes the particular line(s) it is representing to be drawn. Similarly to State 1 the sounds are spatialized to a stereo field so that events occurring on the right side are played through the right channel and events on the left side through the left channel.

The dynamic variety of the computer's response makes the algorithmic sonification somewhat challenging. At times the computer may draw thousands of lines in a span of a few seconds while at other times may draw only one line across ten seconds. Through observation and testing we implemented several hardcoded thresholds so that the audio chooses the appropriate sample based on the circumstances of the drawing. When the features describe a scenario between two thresholds the samples are crossfaded based on the distance the value is from each threshold. For example, when there are less than six lines being drawn simultaneously the system will sonify each one using an appropriate sample from the gesture sound library. As the number of simultaneous lines being drawn rises above six the system continues to sonify the individual lines with gesture sound samples while accompanied by an additional rain stick sample. The rain stick sample volume increases as the individual line sample volumes decrease until a threshold is reached and only the rain stick can be heard. From empirical data we found the sonification to be ineffective when using a unique sound for more than 15 lines simultaneously. For this reason we used a single sample with a dense sonic quality (the rain stick) to represent events in which more than 15 lines were being drawn.

4. CONCLUSION

Drawn Together had a soft opening in February, 2012 and is still a work in progress. There is still more which can be done in order to improve the sonic material. Tweaking current thresholds, adding additional samples to the library, and different processing techniques may produce better results. We hope to explore some of these options and implement them in future installments of the piece.

5. ACKNOWLEDGEMENTS

Drawn Together was developed by a large group of individuals including Marc Downie, Shelley Eshkar, Paul Keiser, Mason Bretan, Jason Clark, Jason Freeman, Ryan Nikolaidis, Gil Weinberg, Tristan Al Haddad, Scot Kittle, Kyan Rahimzadeh, Racel Williams, Claudia Rebola, Pablo Alfaro, Asa Martin.

AQUARIUM FUGUE: INTERACTIVE SONIFICATION FOR CHILDREN AND VISUALLY IMPAIRED AUDIENCE IN INFORMAL LEARNING ENVIRONMENTS

Myounghoon Jeon, Riley J. Winton, Jung-Bin Yim, Carrie M. Bruce, and Bruce N. Walker

Georgia Institute of Technology
Atlanta, GA, USA 30332
{mh.jeon, rjwinton, jyim, carrie.bruce, bruce.walker}@gatech.edu

ABSTRACT

In response to the need for more accessible Informal Learning Environments (ILEs), the Georgia Tech Accessible Aquarium Project has been studying sonification for the use in live exhibit interpretation in aquariums. The present work attempts to add more *interactivity* [1] to the project's existing sonification work, which is expected to lead to more accessible learning opportunities for visitors, particularly people with vision impairments as well as children. In this *interactive sonification* phase, visitors can actively experience an exhibit by using tangible objects to mimic the movement of animals. Sonifications corresponding to the moving tangible objects can be paired with real-time interpretive sonifications produced by the existing Accessible Aquarium system to generate a cooperative fugue. Here, we describe the system configuration, pilot test results, and future works. Implications are discussed in terms of *embodied interaction* and *interactive learning*.

1. INTRODUCTION

To improve the accessibility of exhibits and promote universal design in aquariums, researchers have studied real-time interpretive sonification as a strategy for translating visual aspects of live animal exhibits [2, 3]. Georgia Tech's Accessible Aquarium Project has focused on designing sonifications for individuals with vision impairments that convey the informational (e.g., the number of animals in view, animal locations, and animal movements) and aesthetic aspects (e.g., the "feeling" or mood perceived by visitors) of live exhibits that a visitor might experience when viewing a live exhibit). This enables visitors with vision impairments to experience an exhibit in both cognitive and affective aspects, and it also provides a shared experience so that visitors with and without vision impairments can discuss their understanding and impressions of the exhibit. One way to accomplish this is through music that communicates both information and feeling. Previous studies [2, 3] showed that we could match musical features such as pitch and tempo with animal information such as height in tank and swimming speed to facilitate understanding of exhibit dynamics. The project also has implications such as *biologically inspired music* or *dynamic sonification* from the sonification perspective [3]. To fulfill and strengthen those two aspects, the current project attempts to enrich visitors' experiences in aquariums by combining and harmonizing animal- and audience-inspired sonification. By increasing *interactivity* [1] among animals, people, and sonification systems, it is expected that visitors will have an enhanced learning experience.

2. RELATED WORK

For interactive sonification, embodied interaction has been used and shown effective in various learning and training domains. To illustrate, Antle et al. [4] has used embodied interaction framework to elicit, train, and apply people's embodied metaphors as a means of developing intuitive fluency with music creation. Based on a specific metaphor of "music is physical body movement", they developed a computational system that helps children understand musical concepts such as melody, harmony, and rhythm in the form of intuitive, physical analogs. Howison et al. [5] introduced an embodied-interaction based instructional design, the Mathematical Imagery Trainer (MIT). They aimed at helping young students develop an understanding of proportional equivalence by applying the embodied cognition paradigm, in which mathematical concepts are grounded in mental simulation of dynamic imagery, which is acquired through perceiving, planning, and performing actions with the body. Recently, in the sonification community, several interactive movement projects have been introduced in sports training [e.g., aerobics, 6, rowing in a boat, 7]. All of these projects have suggested that fully engaging embodied interaction with sonified feedback is effective in enhancing the user experience.

3. CONCEPT AND SYSTEM CONFIGURATIONS OF THE CURRENT RESEARCH

In the current research, we attempt to leverage the real-time interpretive sonifications of the Accessible Aquarium Project to enable a collaborative sonification that includes visitor interaction. The real-time interpretive sonification of the exhibit dynamics contains *coherent responses* with consistent *feedback loop*, which is a subset of interactivity [1]. This new work, adds more interactive elements (e.g., *responsiveness*), by allowing visitors to engage with the live exhibits through tangible user interface objects (TUIOs) that represent the animals in the exhibit. Consequently, visitors will contribute to a cooperative sonification of the live exhibit (generating a counter melody). Additionally, it is anticipated that visitors, including those with vision impairments, will learn about animal movement and perhaps, become interested in other interpretive information. For the rapid prototyping, we have taken a simple movement-to-sound mapping approach to complement the real-time interpretive sonification of live animal movement. Figure 1 shows the schematic system configuration. Two cameras can be used for the system: a HD, high speed digital camera to track animals for the real-time interpretive system and a web camera to track the TUIO in the visitor interaction system.

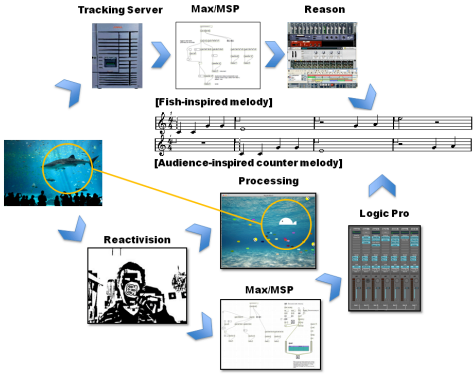


Figure 1: Schematic system configurations of the real-time interpretive sonification and visitor interaction sonification.

While the system is designed to use these two cameras to actively work with real-time computer vision data, the prototype system can also simulate incoming real-time video data by utilizing recorded video. The parameters (x, y coordinates) of visitors' movement is processed in the reacTIVision software and then transmitted into a Max patch that sends MIDI (Musical Instrument Digital Interface) signals into the Logic software which generates the final sounds using virtual instruments.

4. PILOT STUDY

To test our conceptual design, we conducted a pilot study with a prototype "interactive game" scenario via laptop. In this system, users were presented with a short musical motif or melody that was generated based upon the animal movement pattern. The users were then asked to reproduce this melody by maneuvering the TUIOs within the field and select the desired notes with a controller device. Our TUIOs for the prototype were the standard fiducial figures provided by reacTIVision, attached to animal models in order to promote a more concrete link to the virtual animals on the laptop and support interactivity. In this initial prototype, the y-axis of space represented the pitch of the notes (1 octave) and the x-axis represented panning of the sound for spatialization purpose. For example, if the desired sequence was F3-A3-C4-G3-F3, a user would move the TUIO until F3 was heard and then click the controller to select the note. The user would then move the TUIO up to select A3, up for C4, down for G3, and down again for F3, while clicking to select each note as they arrive. This same process would be taken for every note in the sequence and upon correct completion an audible chime and a visual indicator notified the users that they had successfully completed the melody. After an initial training period, we had all participants complete this task using only auditory information, thus providing a simulation of vision impairment as well as focusing the user on pitch detection ("musical ear"). The children (2 female, 2 male, mean age = 5.5) who tested the prototype described the system as, "fun," "interesting," and "engaging." The pilot study yielded several ideas for experimenting with different interface configurations and mappings.

5. DISCUSSION AND FUTURE WORKS

It is important to develop auditory displays that effectively convey exhibit information and aesthetics in order to enhance learning experiences for all visitors, including those with vision impairments. In this work, we are suggesting that visitors can go beyond the limited role of passive learners and explore a more constructive and interactive role. Chi [8] recently provided a framework that offers a way to differentiate *active*, *constructive*, and *interactive* in terms of observable activities and underlying learning processes. Active learning is doing something physically, such as look and fixate. Constructive learning is producing outputs, such as self-explain and elaborate. Interactive learning includes dialogue containing guided-construction, such as revise errors from feedback and co-construction. While active learning is attending processes, constructive learning is creating processes. Interactive learning means jointly creating processes incorporating a partner's contributions. According to Chi, interactive activities are most likely to be better than constructive activities, which in turn might be better than active activities, which are better than being passive. Based on Chi's argument, we are attempting to incorporate interactive activities in this multimodal learning environment, by allowing visitors to interact with animals or peers and construct their own music. To this end, they can create 3rd and 4th counter-melodies by interacting with their fellow visitors. We can employ diverse strategies for using music as sonification. For example, we will investigate various mappings to identify how visitors' approximate movements can create more musically matched sounds. We also plan to integrate two separate sonification systems so that an animal-inspired melody can evolve and adapt to the visitors' music pattern. Furthermore, we expect to incorporate narration as sonification to provide a more transparent form of information and aesthetics. These narrations (or lyrics) could provide verbal descriptions of animal characteristics and facts to accompany the music. These multifaceted efforts are expected to create innovative and engaging soundscapes in aquariums that attract and welcome a wide range of visitors, including those with vision impairments as well as children.

6. REFERENCES

[1] S. Rafaeli, "Interactivity: From new media to communication " Sage Annual Review of Communication Research: Advancing Communication Science, vol. 16, pp. 110-134, Sage: Beverly Hills, CA, 1988.

[2] B. N. Walker et al., "Aquarium sonification: Soundscapes for accessible dynamic informal learning environments.," in International Conference on Auditory Display, London, UK, 2006, pp. 238-241.

[3] B. N. Walker et al., "Musical soundscapes for an accessible aquarium: Bringing dynamic exhibits to the visually impaired," in Proceedings of ICMC, Copenhagen, Denmark, 2007.

[4] A. N. Antle et al., "Playing with the sound maker: Do embodied metaphors help children learn?," in International Conference on Interaction Design and Children, Chicago, IL, 2008, pp. 178-185.

[5] M. Howison et al., "The mathematical imagery trainer: From embodied interaction to conceptual learning," in SIGCHI Conference on Human Factors in Computing Systems, Candada, 2011, pp. 1989-1998.

[6] T. Hermann and S. Zehe, "Sonified aerobics: Interactive sonification of coordinated body movements," in International Conference on Auditory Display, Budapest, Hungary, 2011.

[7] N. Schaffert et al., "The sound of rowing stroke cycles as acoustic feedback," in International Conference on Auditory Display, Budapest, Hungary, 2011.

[8] M. T. H. Chi, "Active-constructive-interactive: A conceptual framework for differentiating learning activities," Topics in Cognitive Science, vol. 1, pp. 73-105, 2009.

BEYOND VISUALIZATION: ON USING SONIFICATION METHODS TO MAKE BUSINESS PROCESSES MORE ACCESSIBLE TO USERS

Tobias Hildebrandt[1], *Simone Kriglstein*[2] *and Stefanie Rinderle-Ma*[1]

[1] University of Vienna, Austria, Faculty of Computer Science
[2] SBA Research, Vienna, Austria
[1] {tobias.hildebrandt, stefanie.rinderle-ma}@univie.ac.at, [2] SKriglstein@sba-research.at

1. INTRODUCTION

Making business processes accessible to users constitutes a crucial challenge throughout their entire life cycle: users should be enabled to understand business process models (*Analysis & Design* phase), keep an overview on running process instances (*Operation* phase), perceive process adaptations (*Operation* phase), and comprehend as well as interpret results of analyzing processes (*Evaluation* phase). What sounds easy for small process models quickly becomes an enormous challenge in the context of complex *wallpaper* process models because they can consist of hundreds of process activities, data flows, and resources and can have thousands of running process instances in different execution states. Obviously, for such scenarios it becomes very hard to recognize or even understand, e.g, deviations from the regular process execution path.

Research has been conducted to analyze how visualization methods can help users to understand processes. There exist several tools that offer process visualization approaches to support users to model and monitor business process models and instance data. However, visualization methods for business processes show several limitations [1]: (a) limited screen size, (b) irregular process patterns, (c) executions or large number of process instances in different execution states, (d) displaying process change information as well as assessing certain process analysis and mining results are difficult, yet crucial. In such cases, it can be beneficial to use data sonification in order to enhance process visualizations. Although many reasons appear to apply sonification for representing process-related data, only very few approaches addressed this issue so far. Kramer et al. [2] found out, that the auditory perception is especially sensitive to temporal change. Furthermore, sonification, in contrast to static visualization, can only exist in time. As process instances per definition can only exist in time as well, sonification naturally lends itself to this area (as do animated visualizations). This promises advances when trying to convey process exceptions and changes to users.

2. SONIFICATION OF BUSINESS PROCESSES

One of the few applications of sonification in the area of business processes is the project Grooving Factory [3] of the Jacobs University Bremen. It aimed to reveal bottlenecks in industrial productions and to improve the logistics by sonifying production processes. The developed prototypes enable users to select the different working stations and manufacturing orders of the production process to be sonified.

In the ARKOLA simulation Gaver et al. [4] describe a live multi-modal sonification of a bottling plant. In this simulation, users manually control the settings and adjustments of several interconnected machines, trying to avoid stops and bottlenecks. Occurring events such as the spill of liquid are being communicated to the user by appropriate sounds.

Besides Grooving Factory and the ARKOLA simulation, there seems to be no research project that deals with the sonification of business processes. Research such as that of Hermann et al. [5] deals with the sonification of processes, but not in corporate or business environments (the mentioned example deals with processes in the area of robotics). This leads to the assumption that there still is a substantial amount of untapped research potential in this area.

In order to answer the question which sonification techniques might be best suited to convey business process information, it seems logical to start with analyzing the type and structure of data that typically accumulates during the individual life cycle phases of business processes and subsequently evaluate accepted sonification techniques in terms of their suitability to convey this process information. Most data in the process design phase is related to static process models and their change history. During process operation, the data that typically accumulates can be grouped into two categories: on the one hand, users want to monitor high-level data that accumulates during the execution of the individual process instances (like the number of running instances per process model, current capacity utilizations or the general *health* of the system). This is quantitative data that is updated in regular intervals. On the other hand, users want to inspect individual process instances in more detail in cases of irregularities or specific situations. This instance data is often not very complex and individual data sets typically consist of event occurrences and a few related data elements (like the name of an activity that has been started or completed, together with the name of the associated user and a time stamp), in some cases coupled with quantitative data. However, the data of one such process instance can, in some cases, consist of thousands of such individual events. During the process analysis phase, users want to analyze this execution data in a retroactive, more condensed manner.

The five most accepted sonification techniques are probably *audification*, *auditory icons*, *earcons*, *parameter-mapping* and *model-based sonification*. The techniques audification and model-based sonification may not seem to be the most obvious choices for the sonification of business-process related data. Audification relies on a huge number of quantitative data, which typically is not available to such an extend in this domain. Additionally, it might be very difficult or even impossible to distinguish between several *streams* of sounds using audification techniques. Concerning model-based sonification, Hermann [6] states that audification or

parameter mapping should be preferred to model-based sonification in most cases in which the data that needs be sonified is time indexed. Data that accumulates in the area of business processes is indeed in large part time-indexed.

Auditory icons have already been applied to sonify static models (e.g., [7]), which suggest that they might be applied during the process design phase to sonify process models. During the sonification of the execution of individual process instances (in the phases operation and analysis), the sonic pendants of the involved activities and events could be played back upon their incidences. As an example, a process event "customer has payed his invoice" could be conveyed by playing the sound of a cash register being opened. Analogous, the sound of a shopkeepers bell could signify the acquisition of a new customer. Depending on the industry and the type of processes, there is often a variety of self-explanatory sounds that can be used in order to sonify the respective events and activities. Thus, it would be possible to recognize deviances of individual process instances from more typical process executions by the fact that the respective sounds are being played in a different order, or in a different *rhythm*.

Earcons are in a similar fashion suitable for sonifications during the life cycle phases design, operation and analysis, but more flexible. For some process events it could prove difficult to find real-world-sonic analogies. For example, it could be a challenge to find sounds that are sonic analogies to the states "customer is already registered" and "new customer". This differentiation would therefore be hard to convey using auditory icons, so the usage of earcons might solve that problem (even if studies suggest that earcons are harder to recognize than auditory icons). By using parameterized auditory icons or earcons, not only the information can be conveyed that a certain event has occurred, but also one or several quantitative data attributes that are connected to that event. One could for example imagine an auditory icon conveying the occurrence of an event "incoming payment", while the sum of the payment is mapped to the pitch of that auditory icon.

Parameter mapping might not be suitable for sonifications in the process design phase, as there is little quantitative data to be mapped, but merely static process models. However, during the process operation phase, parameter-mapping sonifications might be used to map high-level data that accumulates during process executions to one or several sound streams. These sound streams might then be played back continuously which should make it feasible for the user to recognize patterns and modifications as well as to get an overview of the general "health" of individual running processes or a complete system. The same (or similar) concepts might be applied to analyze historic process execution data retroactively.

This extended abstract however constitutes just a preliminary analysis of which sonification techniques might be suitable to support users in their business-process related tasks. A more thorough analysis of the specific characteristics of process-related data in the individual life cycle phases will be necessary before making decisions concerning which sonification techniques will be applied during the development of respective prototypes.

Besides the fact, that different sonification techniques might be adequate for different tasks that users perform during the different life cycle phases of business processes, the two modalities visualization and sonification might also be suitable to different extends for these areas. In the process design phase, visualization might be more suitable than sonification. Graphical user interfaces already allow the user-friendly creation of process models, a task that may not benefit substantially from sonification. However, after (or during) the graphical creation of process models, sonification might well be helpful when it comes to simulating process models in order to test them for potential problems (such as deadlocks). During process operation, a sonification could be used to monitor the execution of process instances. One could imagine, depending on the scenario, either a constant real-time sonification of all running process instances, or an *auditory summary* of a certain time period (for example a shortened sonification of the last 24 hours). In such a sonification it should, after a learning phase, be possible to detect deviances or critical situations during the execution of process instances. A multi-modal solution could combine sonification with the possibility to visually explore root causes or other details, once such a situation has been recognized in the sonification. Similar approaches could be applied in the process analysis phase. In general, multi-modal sonifications of business process-related data should consider the strengths and weaknesses of both modalities in order to be able to best assist users in their tasks.

Future work will result in first recommendations on how to apply multi-modal approaches in the context of business processes along their entire life cycle. Subsequently, prototypes that base on those results will be developed and evaluated.

3. REFERENCES

[1] S. Rinderle, R. Bobrik, M. Reichert, and T. Bauer, "Business process visualization - use cases, challenges, solutions." in *Proc. of the 8th Int'l Conf. on Enterprise Information Systems*, 2006, pp. 204–211.

[2] G. Kramer, B. Walker, T. Bonebright, P. Cook, J. Flowers, N. Miner, J. Neuhoff, R. Bargar, S. Barrass, J. Berger, G. Evreinov, W. Fitch, M. Grohn, S. Handel, H. Kaper, H. Levkowitz, S. Lodha, B. Shinn-Cunningham, M. Simoni, and S. Tipei, "Sonification report: Status of the field and research agenda - report prepared for the national science foundation by members of the international community for auditory display," 1999.

[3] K. Windt, M. Iber, and J. Klein, "Grooving factory - bottleneck control in production logistics through auditory display," in *Proc. of the 2010 Int'l Conf. on Auditory Display (ICAD)*, E. Brazil, Ed. Washington, D.C., USA: International Community for Auditory Display, June 2010. [Online]. Available: http://icad.org/Proceedings/2010/WindtIberKlein2010.pdf

[4] W. W. Gaver, R. B. Smith, and T. O'Shea, "Effective sounds in complex systems: the ARKOLA simulation," in *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems: Reaching through technology (CHI'91)*. ACM, 1991, pp. 85–90.

[5] T. Hermann, C. Niehus, and H. Ritter, "Interactive visualization and sonification for monitoring complex processes," in *Proc. of the 2003 Int'l Conf. on Auditory Display (ICAD)*. International Community for Auditory Display, 2003, pp. 247–250.

[6] T. Hermann, "Model-based sonification," in *The Sonification Handbook*. Logos Berlin, Dec. 2011, pp. 399 – 428.

[7] I. Berman, Lewis, "Program comprehension through sonification," Ph.D. dissertation, Durham University, 2011. [Online]. Available: http://etheses.dur.ac.uk/1396/

SHAKING UP EARTH SCIENCE: VISUAL AND AUDITORY REPRESENTATIONS OF EARTHQUAKE INTERACTIONS

Chastity Aiken

Georgia Institute of Technology
School of Earth and Atmospheric Sciences
311 Ferst Drive, Atlanta, GA 30332
chastity.aiken@gatech.edu

Zhigang Peng

Georgia Institute of Technology
School of Earth and Atmospheric Sciences
311 Ferst Drive, Atlanta, GA 30332
zpeng@gatech.edu

David Simpson

IRIS Consortium
1200 New York Avenue, NW Suite 400
Washington, DC 20005
simpson@iris.edu

Andy Michael

U.S. Geological Survey
345 Middlefield Road, Mail Stop 977
Menlo Park, CA 94025
michael@usgs.gov

Debi Kilb

Scripps Institution of Oceanography, UCSD
9500 Gilman Drive, Mail Code: 0225
La Jolla CA, 92093
dkilb@ucsd.edu

Bogdan Enescu

Earthquake Research Department,
National Research Institute of Earth Science
and Disaster Prevention
3-1 Tennodai, Tsukuba, Ibaraki 305-0006, Japan
benescu@bosai.go.jp

David Shelly

U.S. Geological Survey
345 Middlefield Road, Mail Stop 910
Menlo Park, CA 94025
dshelly@usgs.gov

ABSTRACT

One earthquake can influence subsequent earthquakes. To demonstrate such earthquake interactions, seismologists have used in the past "snapshot" static images. Although static images can, by themselves, convey basic visual information about the spatial distribution of earthquakes, adding auditory information could help to provide additional details on the temporal evolution of the earthquake sequences. Recently we have used standard tools like MATLAB and Quick Time Pro to produce animations with time-compressed sounds to demonstrate both immediate aftershocks and remotely triggered tremors related to the 2011 magnitude 9.0 Tohoku-Oki, Japan, earthquake. Here we show our development in this direction that includes multiple parameters of earthquakes and seismic waves to present the concept of earthquake triggering.

INTRODUCTION

Large shallow earthquakes are typically followed by increased seismic activity near the mainshock rupture, which are termed "aftershocks." Sometimes large earthquakes are preceded by elevated seismic activity known as "foreshocks." Large earthquakes can also trigger shallow microearthquakes and deep tectonic tremors located several hundred to thousands of kilometers away [1], [2]. Many of these events are believed to occur in regions of high fluid pressure, including geothermal/ volcanic areas and the deep extension of some plate-bounding faults. Because earthquakes and tremors occur at different depths on major faults, both are important for understanding the underlying physics of earthquake nucleation, as well as overall fault behavior.

Seismologists have mainly used "snapshot" static images to present their results. Adding auditory information to static images could provide more information about the temporal evolution of earthquake sequences that can be easily communicated to a wide audience – seismologists, educators, and the public. This is due to the fact, as noted by [3], that our eye is strong in identifying *structures, surface,* and *steadiness*, while our ear is good at recognizing *time, continuum, remembrance,* and *expectation.* Hence, by combining audio and video components of seismic data and earthquake information, we can more effectively communicate the integral parts of earthquake sequences to a general audience.

Audification of seismic data is not a new concept [4], [5], [3], but has recently been rejuvenated [6], [7]. Recordings of nearby earthquakes and tremors contain frequencies of up to 100 Hz, which are on the lower end of the audible spectrum (20 - 20k Hz) that humans can hear. One of the more simple ways to map earthquake signals into audible range is to play it faster than the true speed (i.e. time compression) [5]. This also reduces the playback time so audiences can hear seismic signals that typically occur over a few hours in a matter of minutes. Recently, we combined images of waveforms and spectrograms with time-compressed sounds to demonstrate both immediate aftershocks and remotely triggered tremors in California related to the 2011 magnitude 9.0 Tohoku-Oki, Japan, earthquake [7]. These animations were generated using standard tools like MATLAB and Quick Time Pro [6] and represent our initial exploration of the use of both visual and auditory elements to convey information about earthquakes and tremors.

CURRENT AND FUTURE WORK

Here we present new audio-video products that present multiple parameters of earthquakes and seismic waves. The first one contains two or more audio channels (i.e. stereophonic sounds) converted from multiple frequency ranges of the same seismic data. We use the broadband seismogram recorded at station PKD in Central California during the 2011 Tohoku-Oki earthquake as an example. In our previous study [7], we used a factor of 100 to speed up the seismogram to audify teleseismic *P* wave (up to 5 Hz) and the locally triggered tremor signals (up to 30 Hz). However, the speed-up factor is still not high enough to bring the long-period *S* and surface waves (periods of around 20 s) into the audible range.

Rather than increasing the speed-up factor more (which would make the duration of the animation too short), we use MATLAB's built-in function *vco* (voltage controlled oscillator) to create a signal from the broadband seismogram that oscillates around an audible center frequency in proportion to the amplitude of low frequency ground motion. The amplitude (volume) of the *vco* tone is modulated by the envelope function of the broadband data. The length of this representation with both frequency and amplitude modulation (FM/AM) is selected to be the same as that for the time-compressed, high-frequency channel. Finally, we save both outputs as a stereo sound (with MATLAB's *wavwrite* function) and combine them with the static images to generate an animation with sound (See examples at [8]). Because this new version contains sound from dual frequencies, we can hear very clear correlations between the long-period *S* and surface waves and the high-frequency triggered tremor signals. In comparison, our previous product only has the high-frequency tremor signals, and hence such correlation can only be seen from the static images.

Other ongoing efforts include: (1) using seismic data from multiple stations and mixing them, or directly using events listed in earthquake catalogs to give spatial position in stereo or 5.1 surround sound – i.e. the listener at the epicenter and the stereo effect positioning the stations or earthquakes, (2) directly mapping earthquake parameters (epicentral locations, depths, and magnitudes) into sound properties (amplitudes, pitches, and durations). A recent example from another group is shown at [9]. Our goal is to produce innovative and simple ways of presenting earthquake waveforms and sequences and to share the results with other researchers and educators as well as the public. These approaches and products not only provide the general public with a fun and engaging way to gain scientific knowledge of earthquake interaction but may also help seismologists better study the phenomenon and decipher the underlying mechanisms.

REFERENCES

[1] Hill, D. P. and S. Prejean (2007), Dynamic triggering, In *H. Kanamori (ed.) V. 4 Earthquake Seismology*, 258-288, Treatise on Geophysics (G. Schubert, ed. in chief), Elsevier, Amsterdam.

[2] Peng, Z. and J. Gomberg (2010), An integrated perspective of the continuum between earthquakes and slow-slip phenomena, *Nature Geoscience*, **3**, 599-607, doi:10.1038/ngeo940.

[3] Dombois F. and G. Eckel (2011), Audification, In *The Sonification Handbook*, ed. T. Hermann, A. Hunt, and J. Neuhoff, 301-324.

[4] Speeth S. D. (1961), Seismometer sounds, *Journal of the Acoustical Society of America*, **33**, 909-916.

[5] Hayward, C. (1994), Listening to the Earth sing, In *Auditory Display: Sonification, Audification, and Auditory Interfaces*, ed. G Kramer, 369-404, Reading, MA: Addison-Wesley.

[6] Kilb, D., Z. Peng, D. Simpson, A. Michael, M. Fisher, and D. Rohrlick (2012), Listen, Watch, Learn: SeisSound Video Products, *Seismological Research Letters*, *83*(**2**), 281-286.

[7] Peng, Z., C. Aiken, D. Kilb, D.R. Shelly, and B. Enescu (2012), Listening to the 2011 Magnitude 9.0 Tohoku-Oki, Japan, Earthquake, *Seismological Research Letters, 83*(**2**), 287-293.

[8] http://geophysics.eas.gatech.edu/people/zpeng/Japan_20110311/#PKD

[9] http://www.youtube.com/watch?v=2a--NC4Nong

CHIRPING STARS

Katharina Vogt, Visda Goudarzi, Robert Höldrich

Institute for Electronic Music and Acoustics,
University of Music and Performing Arts Graz,
Austria
goudarzi@iem.at, vogt@iem.at, hoeldrich@iem.at

ABSTRACT

Chirping Stars is a tape piece made of the sonification of Twitter data. A snapshot of the popularity of musicians, randomly drawn in March 2012, yielded eight of the most popular stars at that time. Data of their Twitter followers shows the involvement of rapidly evolving fans of the artists on social media. The sonic interpretation of this development is created by mapping the data to parameters that modulate and re-synthesize the sound tracks of the artists.

1. INTRODUCTION

The social networking service Twitter was launched in July 2006 and has currently over 300 million users, generating over 300 million tweets and handling over 1.6 billion search queries per day [1]. Twitter has become increasingly important for the marketing of musicians. The number of tweets containing the artist's or band's name or the number of their followers on Twitter give an accurate and immediate number for their popularity – events like the release of a new album or the naming of a new-born celebrity baby are responded instantly by the Twitter community.

MusicMetric [2] is one online platform providing network data on musics, e.g., time series of total downloads or fans, taken from different social networks (Twitter, Facebook, MySpace, Soundcloud and others). The data of 1000 artists can be accessed via API commands. The provided data allow to analyze different aspects of the fast moving music market.

We chose to display the development of eight artists, that were among the 'top ten' (or rather, top 13, see below) artists worldwide in March 13th, 2012, and see how their career developed as reflected by Twitter. To find the actual music trends, we used the suggested API of Twitter Music Trends [3]. Pieces of the artists themselves, their major hits on YouTube, are processed depending on the data of artist's development. The resulting sound changes the simple music pieces into a noisy world-wide radio show.

2. DATA COLLECTION

2.1. Artists

We found the following eight artists within the top-13[1] of March 13th, 2012:

Bruno Mars (TwitterID 100220864, @brunomars): Twitter fan data available from: Sat Nov 28 2009 01:00:00 GMT+0100 (CET)

[1]The requested data was not available for the top-10, therefore some artists have been left out.

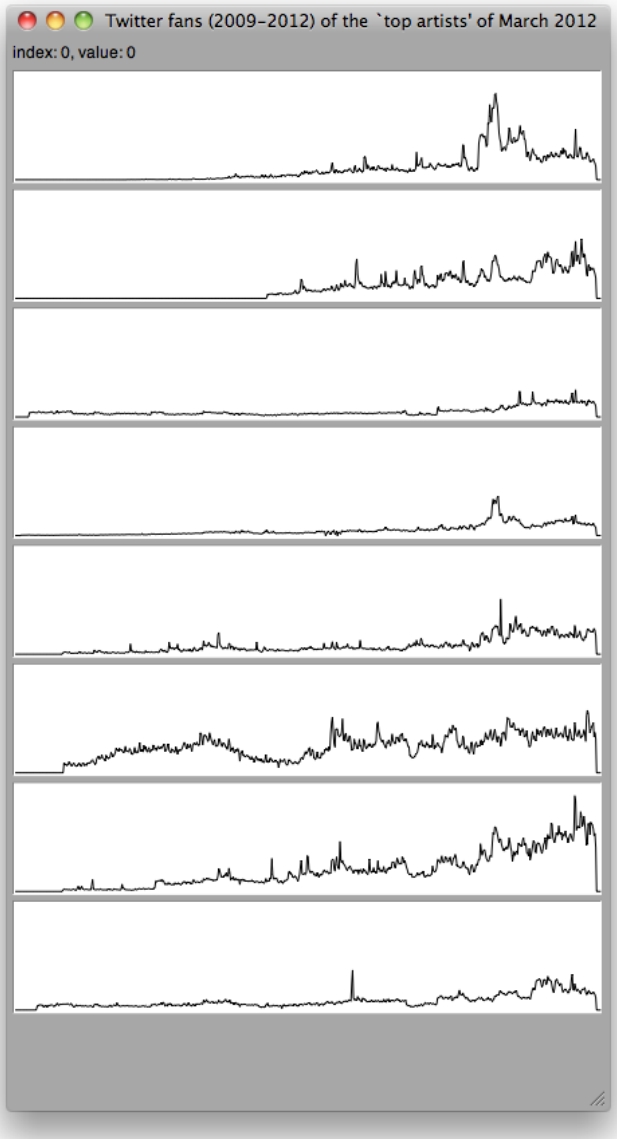


Figure 1: Number of fans on Twitter (followers) for eight artists/ bands until 13th of March 2012. (The sudden onset/offset at the beginning of the data set is due to readout issues and not part of the original data sets.)

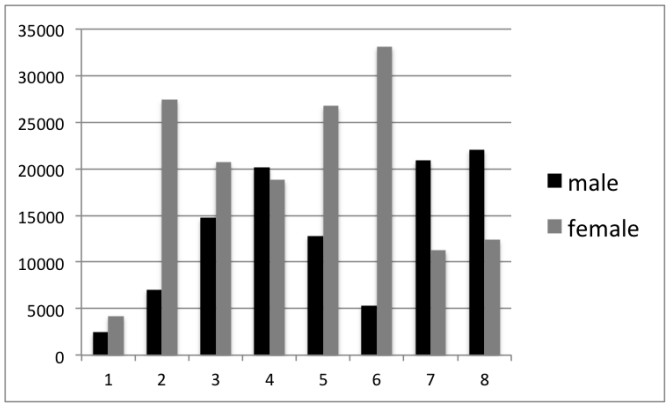


Figure 2: Number of male and female fans on Myspace of the eight artists.

Chris Brown (TwitterID 119509520, @chrisbrown): Twitter fan data available from: Tue Aug 11 2009 02:00:00 GMT+0200 (CEST)

Coldplay (TwitterID 18863815, @coldplay): Twitter fan data available from: Tue Aug 11 2009 02:00:00 GMT+0200 (CEST)

David Guetta (TwitterID 23976386, @davidguetta): Twitter fan data available from: Tue Aug 11 2009 02:00:00 GMT+0200 (CEST)

Drake (TwitterID 27195114, @Drake): Twitter fan data available from: Sat Sep 19 2009 02:00:00 GMT+0200 (CEST)

Justin Bieber (TwitterID 27260086, @Justinbieber): Twitter fan data available from: Wed Nov 25 2009 01:00:00 GMT+0100 (CET)

Rihanna (TwitterID 79293791, @rihanna): Twitter fan data available from: Tue Aug 11 2009 02:00:00 GMT+0200 (CEST)

Snoop Dogg (TwitterID 3004231, @snoopdogg): Twitter fan data available from: Tue Dec 15 2009 01:00:00 GMT+0100 (CET)

Omitted artists (due to missing data) among the top 13: Talor Swift (@Talorswift: 30391175), Lil Wayne (@Lilwayne: 244337185), Adult (@adult : 6224272), Kiss (@kissonline: 22549812) and Wiz Khalifa (@realwizkhalifa: 20322929).

2.2. Data from MusicMetric

We chose to download three different data sets of these artists from MusicMetric: First, the fans (i.e., followers) of the artists on Twitter. Fig. 1 shows the time series of the above artists/ bands from their appearance (some time in 2009) to 13th of March 2012, for a total of 875 days. Different patterns can be found. Many of the stars of that day only became prominent during the last year or so. Sudden outbursts reflect concert dates, CD releases or similar. Due to the chosen data set, the whole sound scene develops to a climax of the presence time.

Second, we gathered information about the gender of the fans, collected as given in specific Myspace fan profiles (see Fig.2).

Third, the location of the fans, collected as downloads of bittorrent files from cities all over the world (see Fig.3).

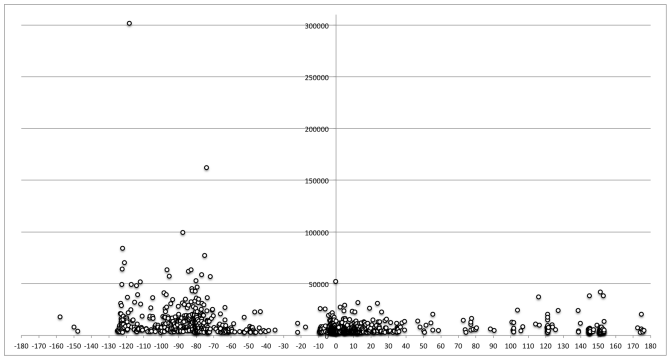


Figure 3: An example of download activity of files associated with *SnoopDogg*, depicted as number of downloads at a certain longitude between -180 and 180 degrees.

3. GLOBAL TWITTER RADIO SHOW

The basic analogy is a global twitter radio show. The sounds are distributed around the listener, who is situated at the Greenwich meridian. From the location data, we use the position of downloads for playing the sound files from various locations, allowing the listener to explore the mainstream music taste in different regions of the world. Gender information is used to filter the sound.

The radio show plays the songs of the above stars in parallel, depending on the data sets. Two sonification approaches were persued. First, granular synthesis serves as a tool where the sound can be well shaped to produce thumbnail patterns and reveal specific information on the data.

Using simple granulation techniques on the data, we obtain a variety of interesting timbres and textures. Gap size, grain size, amplitude, and the random spread of the grains are controlled by the data. Where the number of followers are higher, the samples are played back as recognizable parts of the songs.

Second, we use the data to modulate the original sound file of the artist's song, changing the rate of reading out the sound file which is thus distorted. Here, also the pitch of the tracks is changed, creating a playful, sometimes ironic sound out of the tracks.

4. ACKNOWLEDGMENT

We thank D. Pirrò for his SC3 advice and the Austrian Science Fund FWF for supporting the project *SysSon - A systematic procedure to develop sonifications*.

5. REFERENCES

[1] http://en.wikipedia.org/wiki/Twitter, access: 12/03/15

[2] http://developer.musicmetric.com, access: 12/03/15

[3] http://twittermusictrends.com, access: 12/03/13

THE SOUNDS OF THE DISCUSSION OF SOUNDS

Matt Bethancourt
CUNY Hostos
Media Design Programs
450 Grand Concourse, C Building, Bronx, New York 10451
matt@mouseandthebillionaire.com

ABSTRACT

The Sound of the Discussion of Sounds is a real-time composition and performance computer program built in PHP, MySQL, Max/MSP and Ableton Live that allows listeners to musically hear the subtle changes that occur over time for trending and popular musical artists on Twitter.

1. INTRODUCTION

This is a piece of music that explores the discussion of music. Public opinion is often describes as fickle, but how fleeting is this popularity? Graphs, charts, and datasets give us a glimpse into the short lived nature of fame, but for the most part are limited to two dimensions at a fixed point in time. By making music with data gleaned from social media (Twitter, in this instance), the very nature of the ongoing discussion about musicians and their craft is built into and experienced through a piece of music. This piece will last for as long as the artists it is referencing remain in the popular discussion. As artists lose their hold in the public's eye the piece will become increasingly sparse, until finally there is silence. This may take days, weeks or perhaps even months, and during the process we will be able to hear these changes take place in a way that is both beautiful and thought provoking.

2. PARSING THE JSON DATA

The trending and daily JSON data sets provided by Twitter Music Trends are each parsed with two different sets of PHP programs [Appendix 1-4]. As the program is launched, this PHP code scrapes the top fifty trending artists and the top sixteen artists of the day and inserts each set into a MySQL database. This database is then updated every fifteen seconds, computing the recent change in popularity of each artist, the overall change in popularity of the artist since the program was launched, and the total change of all artist popularity.

Each of the top fifty currently trending artists is also assigned an initial number from one to fifty, that will relate to their musical note in the Max/MSP code. Finally, all of the data and various calculations are published into an HTML page that allows Max/MSP to more effectively parse the data (goo.gl/PJuvZ & goo.gl/mcDAe)

3. MUSICAL ALGORITHMS

3.1. Daily Chord

When the Max/MSP application is launched a root chord is created from the top sixteen artists of the day in the key of A major (from MIDI value 57 to 74). The program continually refers to the parsed JSON data, and every time there is a change in the popularity of any artist, it triggers the note to be played via Ableton Live. The popularity change of each artist increases the velocity of their particular note. In this way, the changes in velocity over time creates subtle changes in the tonal quality of the chord. Meanwhile, the program locates the artist with the highest positive change amount since the last chord was played and applies a short frequency modulation to their corresponding note, further highlighting the movement of the discussion in a musical way.

3.2. Trending Lead

Each of the top fifty trending artists is assigned a sequential note from MIDI value 27 to 77 (D#1 to F5). The current top sixteen notes are sounded in order of popularity at eighth note intervals. When the song first begins this is an ascending D# chromatic scale, but as the program continues to update itself, and the popularity rankings of artists change, interesting patterns and unexpected musical phrases emerge. In addition, the frequency of change in popularity for each artist is directly applied to the quality of the sound of their corresponding note, so that the more change occurs over the course of the piece the less pure the note becomes. Through these methods, the nature of the collective conversation about musicians and their craft shapes the music we are hearing.

4. EXAMPLES AND DOCUMENTS

[1] Project Website: http://www.mouseandthebillionaire.com/icad/
[2] MP3 recording of the first 5 minutes of the program running: goo.gl/P1HpP
[3] The Max/MSP Code can be found here: goo.gl/uxmLm.
[4] Ableton Live project can be found here: goo.gl/guPFE
[5] Pertinent PHP code included in Appendix

5. APPENDIX

1. Daily PHP Initial

```
<?php

// PHP connection code omitted

$json = file_get_contents("http://
twittermusictrends.com/daily.json");
$json_a = json_decode($json, true);

// Clear table of any existing information

mysql_query('TRUNCATE TABLE tbl_daily;');

// Insert Artists into Database

$artist = 0;
while ($artist <= 49){
  $artistName = $json_a["daily"][$artist]
["name"];
  $artistName_fixed =
mysql_real_escape_string($artistName);
  $artistMBid = $json_a["daily"][$artist]
["mbid"];
  $artistInitialScore = $json_a["daily"]
[$artist]["score"];
  $query_insertArtists = "INSERT INTO
tbl_daily (artistID, artistName,
artistMBid, artistInitialScore,
artistPreviousScore, artistCurrentScore)
VALUES ('$artist', '$artistName_fixed',
'$artistMBid', '$artistInitialScore',
'$artistInitialScore',
'$artistInitialScore');";
    mysql_query($query_insertArtists) or
die ("Error in query:
$query_insertArtists");
  $artist++;
}?>
```

2. Daily PHP Updater

```
<?php

// PHP connection code omitted

$json = file_get_contents("http://
twittermusictrends.com/daily.json");
$json_a = json_decode($json, true);

// Find new score and insert into database

for ($artist=0; $artist<=49; $artist++){
  // Get the artist
  $mbid = $json_a["daily"][$artist]
["mbid"];
  // Get new scores
  $tempNewScore = $json_a["daily"]
[$artist]["score"];
  $newScore = round($tempNewScore, 6);
  // Get their record
  $query_rs_artistCheck = "SELECT
artistID, `artistName`, `artistMBid`,
artistInitialScore, artistCurrentScore
FROM tbl_daily WHERE artistMBid='".
$json_a["daily"][$artist]["mbid"]."' ORDER
BY artistID ASC";
  $rs_artistCheck =
mysql_query($query_rs_artistCheck,
$conn_icad) or die(mysql_error());
  $row_rs_artistCheck =
mysql_fetch_assoc($rs_artistCheck);
  $totalRows_rs_artistCheck =
mysql_num_rows($rs_artistCheck);
  // Get Previous Score
  $previousScore =
$row_rs_artistCheck['artistCurrentScore'];
  if (isset($previousScore)) {
    $previousScore = $previousScore;
  } else {
    $previousScore = 0;
  }
  // Calculate score difference
  $scoreChange = $newScore -
$previousScore;
    // Assign score direction identifier
  if ($scoreChange == 0) {
    $changeDirection = 0; // none
  } elseif ($scoreChange < 0) {
    $changeDirection = 1; // down
  } else {
    $changeDirection = 2; // up
  }
  // And add them to the database
  $query_updateArtist = "UPDATE tbl_daily
SET artistPreviousScore='".
$previousScore."', artistCurrentScore='".
$newScore."', scoreChangeAmmount='".
$scoreChange."', scoreChangeDirection='".
$changeDirection."' WHERE artistMBid='".
$mbid."'";
  $updateArtist =
mysql_query($query_updateArtist);
}

// Query updated Artists in the Database

$query_rs_artists = "SELECT artistID,
`artistName`, `artistMBid`,
artistInitialScore, artistPreviousScore,
artistCurrentScore, scoreChangeAmmount,
scoreChangeDirection FROM tbl_daily ORDER
BY artistID ASC LIMIT 16";
$rs_artists =
mysql_query($query_rs_artists, $conn_icad)
or die(mysql_error());
$row_rs_artists =
mysql_fetch_assoc($rs_artists);
$totalRows_rs_artists =
mysql_num_rows($rs_artists);
?>

// HTML displaying results omitted
```

3. Trending PHP Initial

```
<?php

// PHP connection code omitted

date_default_timezone_set('America/
New_York');

$json = file_get_contents("http://
twittermusictrends.com/trending.json");
$json_a = json_decode($json, true);

// Clear table of any existing information

mysql_query('TRUNCATE TABLE
tbl_trending;');

// Insert Artists into Database

$artist = 0;
while ($artist <= 49){
  $artistName = $json_a["trending"]
[$artist]["name"];
  $artistName_fixed =
mysql_real_escape_string($artistName);
  $artistMBid = $json_a["trending"]
[$artist]["mbid"];
  $artistInitialScore =
$json_a["trending"][$artist]["score"];
  $query_insertArtists = "INSERT INTO
tbl_trending (artistID, artistName,
artistMBid, artistInitialScore,
artistPreviousScore, artistCurrentScore)
VALUES ('$artist', '$artistName_fixed',
'$artistMBid', '$artistInitialScore',
'$artistInitialScore',
'$artistInitialScore');";
    mysql_query($query_insertArtists) or
die ("Error in query:
$query_insertArtists");
  $artist++;
}?>
```

4. Trending PHP Updater

```
<?php

// PHP connection code omitted

$json = file_get_contents("http://
twittermusictrends.com/trending.json");
$json_a = json_decode($json, true);

// Find new score and insert into database

for ($artist=0; $artist<=49; $artist++){
  // Get the artist
  $mbid = $json_a["trending"][$artist]
["mbid"];
  // Get the new scores
  $newScore = $json_a["trending"][$artist]
["score"];
  // Get their record
  $query_rs_artistCheck = "SELECT
artistID, `artistName`, `artistMBid`,
artistInitialScore, artistCurrentScore
FROM tbl_trending WHERE artistMBid='".
$json_a["trending"][$artist]["mbid"]."'
ORDER BY artistID ASC";
  $rs_artistCheck =
mysql_query($query_rs_artistCheck,
$conn_icad) or die(mysql_error());
  $row_rs_artistCheck =
mysql_fetch_assoc($rs_artistCheck);
  $totalRows_rs_artistCheck =
mysql_num_rows($rs_artistCheck);
  // Get Previous Score
  $previousScore =
$row_rs_artistCheck['artistCurrentScore'];
  // Calculate score difference
  $scoreChange = $newScore -
$previousScore;
  // Assign score direction identifier
  if ($scoreChange < 0) {
    $changeDirection = 0; // none
  } else if ($scoreChange = 0) {
    $changeDirection = 1; // down
  } else {
    $changeDirection = 2; // up
  }
  // And add them to the database
  $query_updateArtist = "UPDATE
tbl_trending SET artistPreviousScore='".
$previousScore."', artistCurrentScore='".
$newScore."', scoreChangeAmmount='".
$scoreChange."', scoreChangeDirection='".
$changeDirection."' WHERE artistMBid='".
$mbid."'";
  $updateArtist =
mysql_query($query_updateArtist);
}

// Query updated Artists in the Database

$query_rs_artists = "SELECT artistID,
`artistName`, `artistMBid`,
artistInitialScore, artistPreviousScore,
artistCurrentScore, scoreChangeAmmount,
scoreChangeDirection FROM tbl_trending
ORDER BY artistCurrentScore DESC LIMIT
50";
$rs_artists =
mysql_query($query_rs_artists, $conn_icad)
or die(mysql_error());
$row_rs_artists =
mysql_fetch_assoc($rs_artists);
$totalRows_rs_artists =
mysql_num_rows($rs_artists);

?>

// HTML displaying results omitted
```

AFFECTIVE STATES: ANALYSIS AND SONIFICATION OF TWITTER MUSIC TRENDS

Kingsley Ash

Leeds Metropolitan University,
Caedmon Hall 220,
Headingley, Leeds, LS6 3QS, United Kingdom
k.ash@leedsmet.ac.uk

ABSTRACT

This paper describes an approach to the sonification of real-time twitter music trend data realized for the ICAD 2012 Sonification Competition: Listening to the World Listening. The paper will discuss the techniques used to create the sonification and the motivations behind them, including details of the data analysis, mapping strategies, visual display and sonification output.

The system analyses the Twitter Music Trends data feed, which aggregates music listening data from Twitter by artist, as well as the Echo Nest REST API to determine the perceived emotional affect and prevailing descriptions of a selection of the latest trending artists. The resulting data is visualized and sonified in real-time to facilitate analysis and generate an appealing visual and auditory display of the resulting data.

Experience with the system suggests that it is successful in allowing users to determine perceived emotional affect and quality for a number of artists simultaneously, and could allow further investigation into the correlation between these factors. The system also generates appealing visual music that reaches beyond the practice of scientific investigation to reach out to a wider audience.

1. INTRODUCTION

The ICAD 2012 sonification competition is inspired by the idea of music (or sound) about listening to music. It is also inspired by the radical changes over the past decade in how we listen to music and how we share our listening activities with others. Adopting the theme 'Listening to the World Listening', it challenges us to explore what we can learn about listening through the analysis and sonification of social media data about listening.

This response to the competition theme explores our relationship with music listening, both through the new media in which we are now able to share and discuss this music (twitter, blogs, online reviews, etc.) as well as through the words used to describe, categorize and discuss music. In particular, the system examines the words we use to describe the emotional affect the music has on us ("this music makes me feel relaxed/bored/uplifted/angry/etc."), and the words we use to subjectively categorise our listening ("this music is good/bad/fast/slow/unique/predictable/etc."), allowing connections and relationships between these areas to be uncovered and experienced. Furthermore, as this analysis is taking place on live data from Twitter, it allows us to view these relationships in real-time, as the artists in question are listened to and discussed on social media.

The system, created in Max/MSP, contains processes for real-time data collection, data analysis, sonification and visualization and all are discussed in detail in this paper.

2. DATA COLLECTION

The data for this sonification system is obtained from the Twitter Music Trends data feed, which returns a list of the top 50 current trending artists on Twitter. The data is updated every two seconds, and each update is acquired by the system in real-time. From this data the system extracts the artist name, score (ranking popularity of artist on Twitter relative to the most popular artist on the list) and Echo Nest id.

Having obtained the latest Twitter trend data, the system then uses the Echo Nest id tags to access additional information about each artist. In order to gather a range of text documents relating to each artist, the Echo Nest REST API is accessed, using bucket terms to return the 15 most recently available news items, blogs and reviews.

Having completed this process, the system can now analyse the current list of trending artists and a selection of articles in which the artist and their work is discussed. This allows the system to determine the perceived emotions and prevailing descriptions of the artists currently being listened to and discussed across the globe.

3. DATA ANALYSIS

3.1. Valence-arousal space

The most widely used model for the representation of emotions is the 2D valence/arousal space, in which valence is represented on the X axis from highly negative to highly positive, and arousal is represented on the y axis, from calming/soothing to excited/agitated [1]. This model has been widely used to determine the apparent mood of music [2,3] and also to form the basis of music recommendation services [4].

The circumplex model of affect places a number of common emotion words in this 2D valence/arousal space, with negative high arousal emotions in the top left, (e.g. anger, tension, etc.), positive high arousal emotions in the top right (e.g. joy, excitement, etc.), negative low arousal emotions in the bottom left (e.g. boredom, depression, etc.) and positive low arousal emotions in the bottom right (e.g. serenity, calmness, etc.) as shown in figure 1.

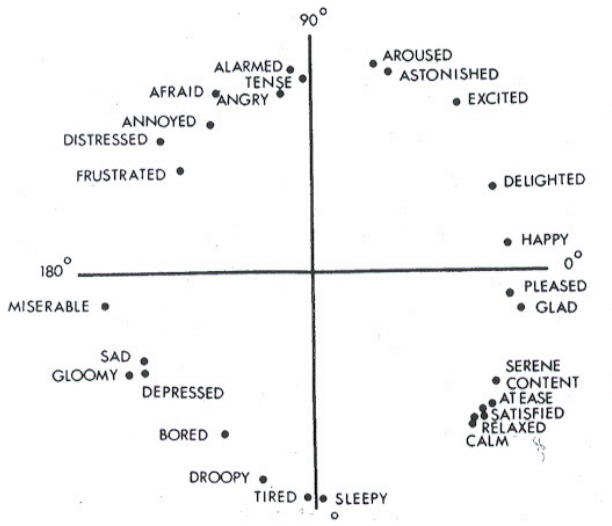


Figure 1: Russell's circumflex model of emotion.

The example emotion words given by Russell and shown in Figure 1 are not always those that come to mind when discussing popular music in the twenty-first century. As a result, a selection of more commonly used emotion words was extracted from a review of online writings referring to popular music artists. Combined with the original emotion words proposed by Russell, a total of 39 words were placed into the valence arousal model (see Figure 2).

High arousal High valence	Low arousal High valence	Low arousal Low valence	High arousal Low valence
Aroused	Pleased	Bored	Confused
Astonished	Glad	Depressed	Frustrated
Stunned	Emotional	Gloomy	Distressed
Excited	Serene	Sad	Annoyed
Inspired	Satisfied	Sorrowful	Painful
Uplifted	Relaxed	Miserable	Aggressive
Surprised	Calm		Angry
Compelled	Soothed		Alarmed
Interested	Dreamy		Tense
Delighted	Sleepy		
Cheerful			
Playful			
Happy			

Figure 2: Emotion words grouped by arousal and valence

For each artist, the system searches through the text obtained from the Echo Nest API to find occurrences of any of these 39 emotion words. Each occurrence is then logged and the total number of each is calculated to give an estimation of the overall current emotional response to each artist.

3.2. Quality-energy space

In order to determine the prevailing descriptions used for each artist, descriptive words are searched for in the same way as above. These descriptive words are also placed in a 2D space, with *quality* running along the X-axis, from bad to good and perceived *energy* along the Y-axis, from slow to fast (see Figure 3).

The number of occurrences of each word is logged, and then the numbers are aggregated for each quadrant of the 2D quality-energy space. These numbers are then combined to give a single measure of quality and a single measure of energy.

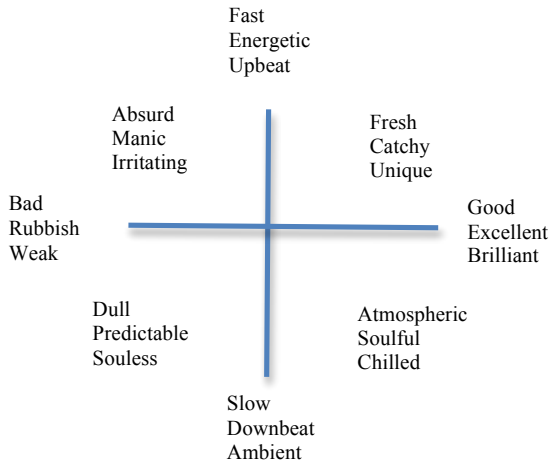


Figure 3: 2D quality-energy space

Having completed the valence-arousal and quality-energy analysis, the results are then used to generate visual and audio outputs for analysis through observation and listening.

4. VISUALISATION

The results of this data analysis are shown in a visualization that simultaneously displays the valence-arousal and quality-energy data for each artist in real-time. A central dot is plotted for each artist in an X-Y space corresponding to their quality-energy score. In this way, the higher the perceived quality of the artist, the further to the right they will appear. Similarly, the higher the perceived energy of an artist, the higher up the visualization space they will appear. As an example, Sigur Ros who are known for their downtempo compositions appear very low down the Y-axis, while the high-energy Japanese pop act AKB48 appear high up the Y-axis.

Having determined the X-Y position of the artist, the emotion words are then plotted around that point at angles corresponding to their position in the 2D valence-arousal model. The length of the lines for each word corresponds to the number of appearances of that word in the text documents. Finally, the name of the artist and their position in the top 50 trending artists is displayed alongside the plot (see Figure 4).

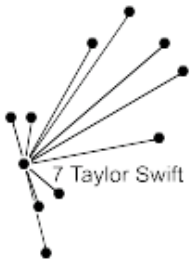


Figure 4: Valence-arousal plot for the artist Taylor Swift.

As can be seen from comparison of Figure 4 and Figure 1, the artist Taylor Swift clearly appears to evoke positive, high arousal emotions such as excitement, though with some elements of positive low arousal states such as calmness and relaxation. The completed plot for each artist therefore shows the arousal-valence scores plotted in an X-Y position that corresponds to the quality-energy scores, allowing all of these parameters to be viewed simultaneously, alongside the artist name and rank.

5. SONIFICATION

Each artist is represented by an individual tone, the timbre of which is synthesized in response to their arousal-valence scores. The fundamental frequency of this sound is determined by the horizontal position on the quality-energy space, and the duration of the sound is determined by the vertical position.

The resulting tone for each artist is synthesised using four components, each corresponding to one quadrant of the valence-arousal model. Positive high arousal emotions (e.g. excitement) are represented using additive synthesis, where the individual scores of each emotion word determine the frequency and amount of the harmonics. Negative high arousal emotions (e.g. annoyance) introduce inharmonic components into the sound, the frequency and amplitude of which are determined by the individual emotion words. Negative low arousal emotions (e.g. tiredness) are represented using filtered noise components in the attack section of the sound, with the centre frequency of each filter determined by the individual scores of each emotion word. Finally, the positive low arousal emotions (e.g. relaxation) are represented by sub-harmonics below the fundamental frequency of the tone. By this method, scores in each quadrant of the valence-arousal space introduce a clearly defined component to the tone that can be picked out in the timbre of the overall sound (see Figure 5).

The overall sound for each artist is played immediately following the collection and analysis of data for that artist. This results in a sonification consisting of a series of consecutive tones allowing both individual and comparative analysis of the data.

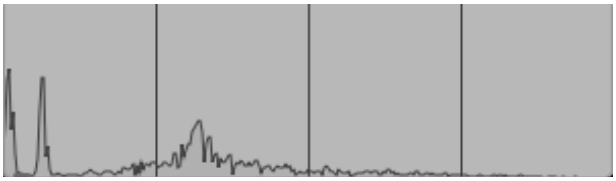


Figure 5: Spectrogram showing fundamental, sub-harmonic and higher frequency noise components of the resulting tone.

6. DISCUSSION

The system successfully enables simultaneous real-time visualization and sonification of music trend data and the results of text-based analysis of recent news, blogs and reviews about each artist. The text analysis process is based on sound foundations in emotional psychology, but there are limitations to the implementation in this system. The analysis looks for individual words, with no regard for the context within which they are used. However, many appearances of a particular word are likely to have some significance, and the combination of a number of words in a similar area of the valence-arousal or quality-energy space are also likely to be significant.

The visualisation is a useful tool to examine the data, allowing both observation of specific points of interest, as well as more general indications of groupings, similarities, trends and movements. However, there is room for further development of this aspect of the system. For example, the visualization doesn't take into account the popularity score of each artist, and this could be utilized as an extra parameter to determine size of the plot for example.

The sonification is also successful in that it allows the listener, with practice, to monitor a large number of simultaneous parameters of the data. The sequential sonification of the data produces repeating patterns in which small variations are easy to detect, but which may turn out to lack interest during repeated listening.

Overall, the system appears to achieve its aim of allowing users to determine perceived emotional affect and quality for a number of artists simultaneously, in a visually and aurally appealing way. Further experience with the system is required to fine-tune the sonification and visualisation algorithms, as well as to fully understand its possibilities and limitations.

7. REFERENCES

[1] J. A. Russell, "A Circumplex Model of Affect," *J. Personality and Social Psychology,* vol. 39, no. 6, pp. 1161-1178, 1980.

[2] D. Lu, L. Liu, and H. Zhang, "Automatic mood detection and tracking of music audio signals," in *IEEE Transactions on Audio, Speech and Language Processing,* vol. 14, no. 1, pp. 5-18, January 2006.

[3] C. Laurier, M. Sordo, J. Serra, and P Herrera, "Music mood representations from social tags," in *Proc. of Int. Soc. For Music Information Retrieval Conf. (ISMIR)*, Kobe, Japan, 2009.

[4] J. Kim, S. Lee, S. Kim, and W. Yoo, "Music mood classification model based on arousal-valence values," in *Proc. Of Conf. of Advanced Communication Technology (ICACT),* Seoul, Korea, 2011, pp. 292-295.

SONIC WINDOW #1 [2011] – A Real Time Sonification

Andrea Vigani

Como Conservatory
Electronic Music Composition Department
Italy
anvig@libero.it

ABSTRACT

This is a real time audio installation in Max/MSP. It is a sonification of an abstract process: the writing on Twitter about music listening experiences on the web from people around the world. My purpose is not to sonify the effects of this process on a musical structure of the songs listened to, like a real-time-echo-web-mix or a new version of J. Cage "Imaginary landscape n°4", but to sonify the structure of the process itself, with its language transducers, its media and its rules. For this purpose, I created a musical instrument played by the data, like a wind chime, but here all the sounds are created by the web data itself, as if the material of a wind chime were the wind itself. It's like an open window on the web listeners where you can observe the action of listening and talking about music, but you don't hear the music listened to and you search for connections, reactions, interactions among the listeners, the transmission media and the code language.

1. DATA USED

Social Genius has created a web service: Twitter Music Trends, which listens to a vast selection of music-related tweets, and automatically tries to detect if each, at that moment, is discussing as a single musician or as a group http://twittermusictrends.com/latest.json (updated every 2 seconds). Information about Twitter music data and the latest artists can be identified from the Twitter stream and the latest 10 IDs of associated tweets.

2. LISTENERS - WRITERS

First of all, the listening process and the tweet process from twitter users; people listen to music and then write tweets about it: it's a human thought about listening to music expressed in a verbal language and syntax. People think, listen and interact with the process and the media with a GUI that translates an information flux. This translation is from a human thought(with its specific language and syntax) to a universal ASCII number code or numeric streams; characters are the same, but syntax changes (ASCII numbers are the common atoms [letters] among different languages) according to an internet code data: **language and syntax change, but information doesn't change.** (Figure 1.)

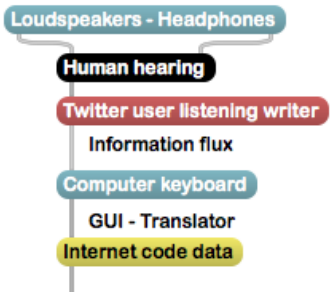


Figure 1.

3. INTERNET CODE DATA ANALYSIS

At this point of the process (that I want to sonify), there is a transduction of the language: the code data from twitter is analysed and the information flux changes: **language and syntax (code) are the same, but information changes:** information is about the process itself, not the original information thought and posted on the web by the twitter users, but a new thought about the first action: the **new information is always a consequence of the previous thoughts.** (Figure 2.)

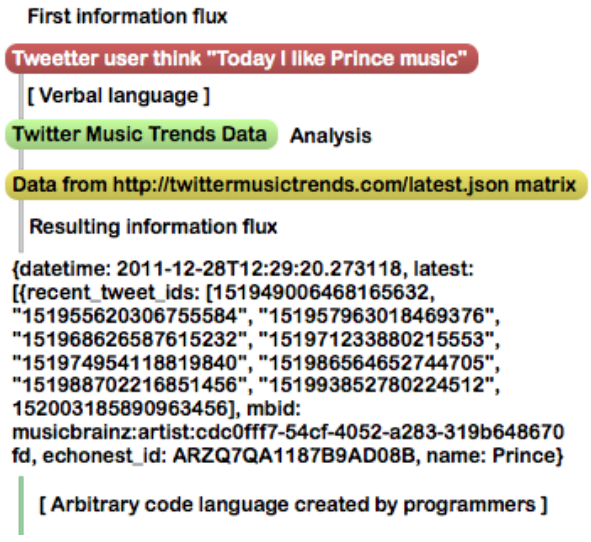


Figure 2.

4. INFORMATION USED

For this sonification I used only one kind of information: NAMES: 1) the Artist Name ; 2) the last 10 Twitter IDs that wrote about the artist (names translation in a code language). In this way, I have a list of 11 names in two different languages (spoken and codified) and these names are connected by a common thought in different ways: the 10 ID names write about the musical actions created by the artist name: names change but the process is always the same, like the musical language…these data becomes in different ways the sound itself and also the score.

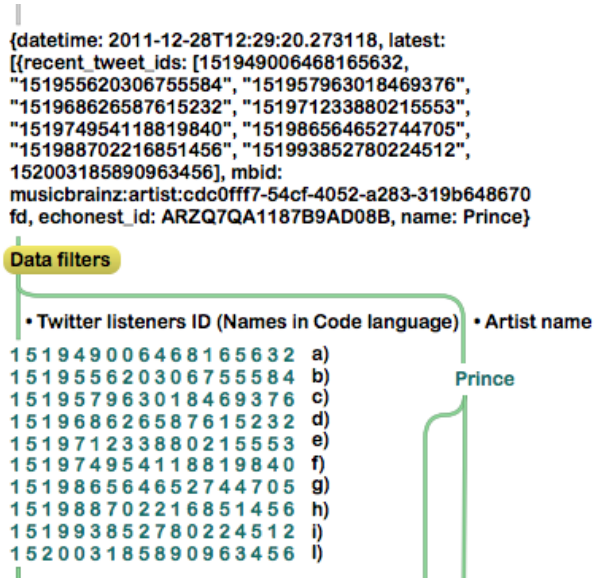


Figure 3.

5. WAVETABLE PLAYER – BACKGROUND NOISE

I used the "last" ten ID numbers scaled from -1 to 1 as amplitudes of a wave-table *(each ID = 18 numbers =180 numbers * 5 (downsampling factor of 2) = 900 samples stored in the wave table) (*Figure 4). They are updated every 2 seconds, according to a choice of the Social Genius programmers and so I programmed a linear interpolation of ID values between the updated triggers, to simulate that the process is continuous.

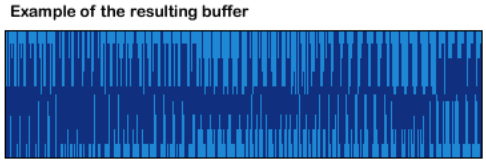


Figure 4.

The wave-table is then played back in a loop at a frequency that varies cyclically from 0.1 to 1.5 Hz, and it's a musical representation of the twitter code web rhythm (a background noise from a portion of the web) morphed by the twitter users almost in real time. At the end of the process, I use a cyclic stereo pan and a cyclic fade-in fade-out to give more sense of "web data waves", as if the web data were a living entity with its own cycles of life.

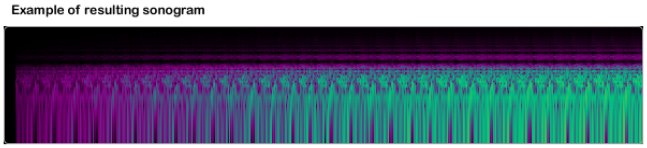


Figure 5: Listen to audio file "1-Background_noise.mp3"

6. SPEECH SYSTEM PLAYER

I use the Artist Name data in two different ways:

1) The Artist Name is translated by the Speech computer software (at each new name, the voice, which reads the name, changes randomly, depending on the computer speech software); then the speech signal passes into a granular synthesis module with a buffer of 10 seconds:

Twitter IDs control in real time:
• grain duration (Min/Max),
• rests between grains ((Min/Max-Voice numbers),
• grain amplitudes and
• grain pan-pot (MIDI)

In this way, the multitude of twitter users voices listening to the artists and also the translation process are represented; at the beginning of the process, the spoken words are translated in ASCII numbers and these numbers are the code "letters-phonenmes"; at that point, with a granular synthesis, I deconstruct the spoken languages (English, French, Italian, etc.) into phonemes (musical language).

Language conversions:

• thoughts (spoken language)→Words written on keyboard→ASCII code→ web code data
• web code data→ASCII code→Spoken language →Phonemes (musical language)

2) The previously obtained "twitter ID background noise" is then filtered by the "last artist name", as if the name could sculpt its profile in the noise: the noise passes into a bank with a maximum of 18 pass filters and frequencies of each filter are given by a conversion of ASCII numbers in frequencies.

Example:

Beatles =
66 101 97 116 108 101 115 (ASCII-Midi Pitches) =
369 2793 2217 6644 4186 2793 6271 Hz (Filter bank center frequencies)
The bandwidths of the filters are given by one of the twitter IDs (scaled from 0.1 to 4 Hz) that is listening to the Beatles:

Twitter IDs: 1 5 0 0 9 6 8 5 4 9 0 0 6 7 8 6 5 6
Bandwidths: 0.8 2.4 0.4 0.4 4. 2.8 3.6 2.4 2. 4. 0.4 0.4 2.8 3.2 3.6 2.8 2.4 2.8 Hz

Each Artist Name is updated every 2 seconds, so the timbre changes without an interpolation every 2 seconds like a "bell signal" and gives a regular beat to the time.

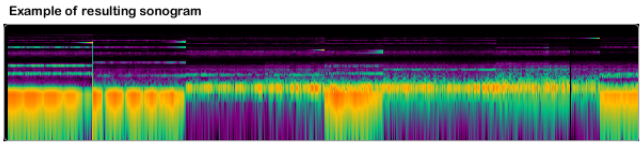


Figure 6: Listen to audio file "2-Timbre_name.mp3"
Listen to audio file "3-Voice_grain.mp3"

7. DATA GLITCHES

One of the last ID listeners gives a small amount of samples stored in a wave- table and played immediately; the amplitudes, which are not scaled and are from 0 to 9 , are afterwards clipped to 1 (wave-shaping) with a linear interpolation between samples. Then the signal is passed through a resonant bandpass filter with a central frequency set to 2000 Hz, bandwidth of 23 Hz and a resonant factor of 3; this gives a "percussive mallet" sound. A quartic envelope is applied to the signal, which has been extracted from the artist name, and the resulting signal enters in a variable delay with a feedback of 1%. This because "the latest artist" scrolls back in position on time... and 2 seconds later he is not ' the latest one' but it's always listened to on twitter; in this case, it doesn't disappear but becomes like an "aura", which gives this sense of slow down and fading, passing through a granular synthesis.

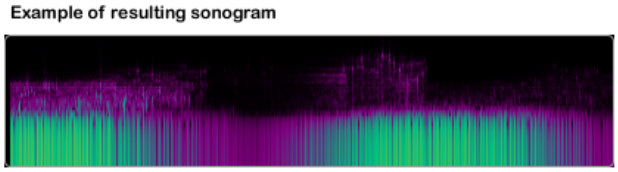


Figure 7: Listen to audio file "4-Data_glitches.mp3"

8. SINE WAVES OSCILLATOR BANK

The last sound generator is an additive synthesis with 18 partials (the number of numbers in a single Twitter ID ; 5 Twitter IDs are mapped according to:

- Frequencies of each partials
- Detuning factor of each partials
- Relative amplitudes of each partials
- Relative durations of each partials
- Relative attack times of each partials

As the IDS are from different people, I applied a granular synthesis to simulate the contemporary presence of 5 different people (the Ids), that are producing the same sound together.

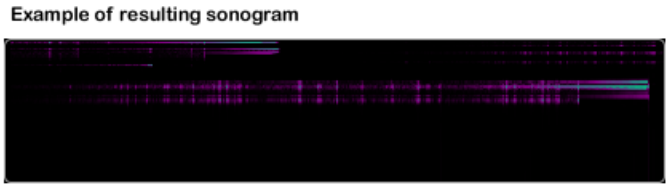


Figure 8: Listen to audio file "5-Oscilbank.mp3"

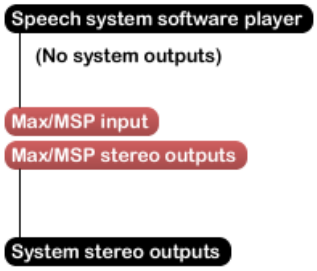
9. EQUIPMENTS AND DIFFUSION

1 Apple computer
1 Internet connection
1 or more Headphones or
1 Audio cart
1 Mixer console table
from 2 to 32 Loudspeakers

It is possible to listen to this audio installation from different computers and headphones or to diffuse the sound on several loudspeakers, to obtain a double interaction: on the other side of the web the listeners create the sounds and on this side other people diffuse this sound in a room and it may be that twitter users, who are present in the room, can change the sound itself…

10. TECHNICAL DETAILS

This software is a Max/MSP patch and you can launch it as an alone application or inside Max/MSP, according to externals used in the patch until now; it is possible to run it only on Apple computers. If you listen to it directly from your computer audio device, it is necessary to do an internal routing; in fact, audio from speech system player will not diffuse out directly, but only after being processed by Max/MSP.



It is possible to route it internally with the software "Sound flower" (from Cycling74 or "Jack") or externally with a sound card, which is present in the room and can change the sound itself…

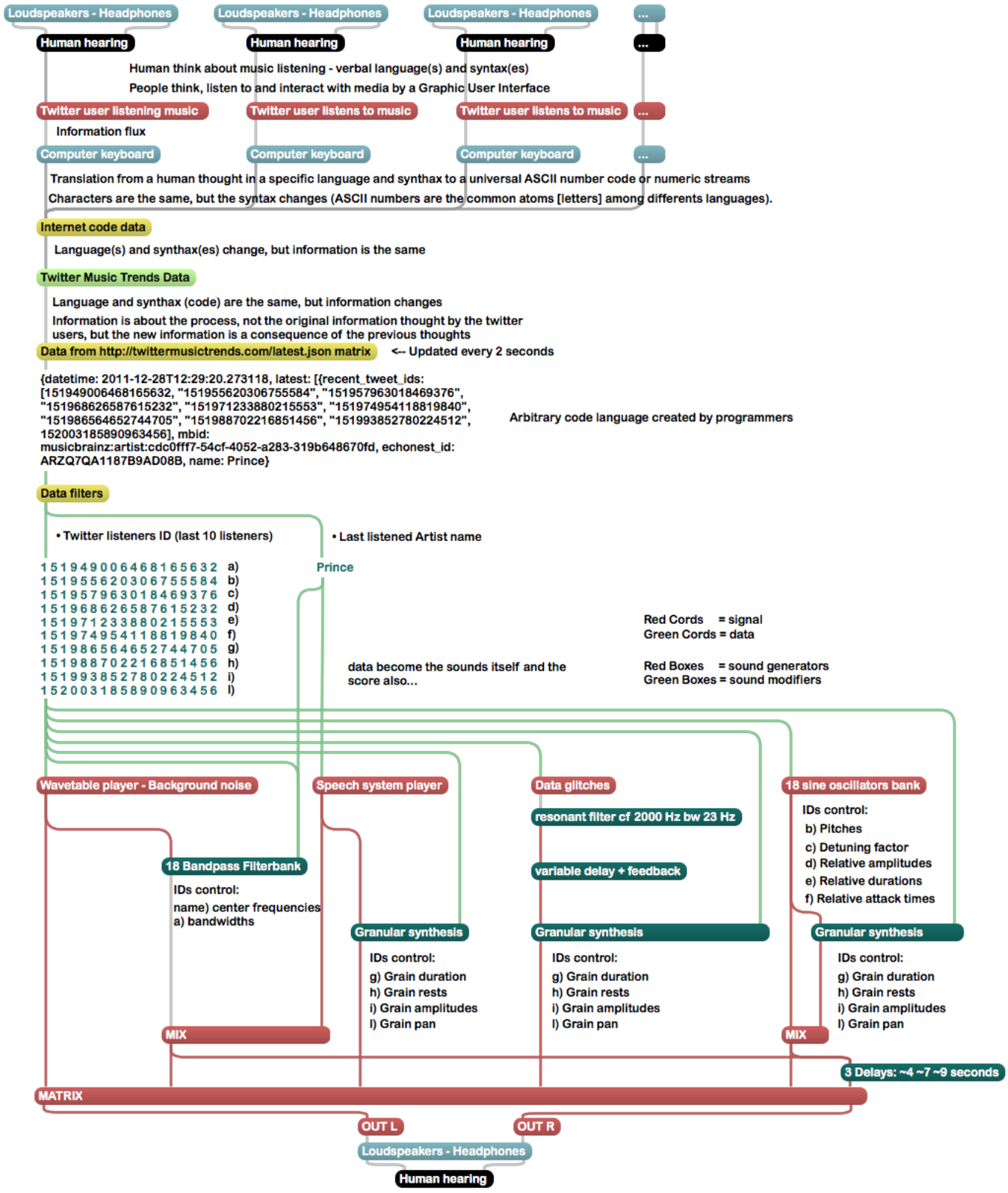


Figure 9: Main Block Diagram