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

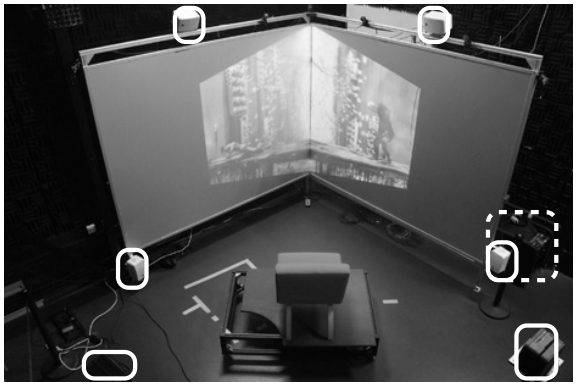

Figure 1: Photo of the SMART-I^2 installation for cinema projection highlighting the image correction for perceived plane projection at the experimental viewing position. —: Surround speakers, - -: Subwoofer.

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`

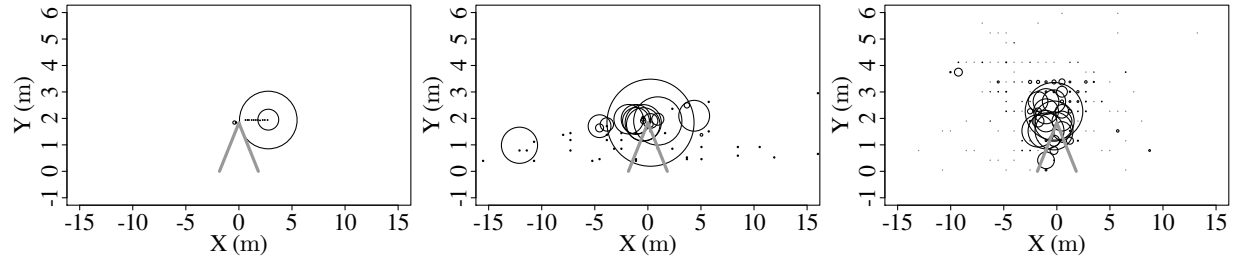

Figure 2: Bubbleplots of the sound source positions in the horizontal plane, taken every 5 frames for the first (left), second (center), and third (right) scenes. Diameters are proportional to the density of sources at that position. Some very distant source positions are not shown for clarity. Note: 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.

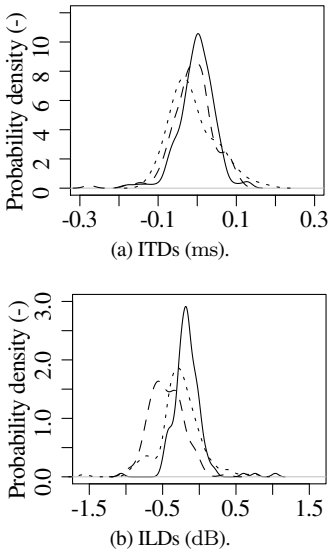

Figure 3: Estimates of the probability density functions of the mean interaural time differences (ITDs) and interaural level differences (ILDs) obtained for each soundtrack. – `ST`, - - `HYB`, ··· `WFS`.

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.

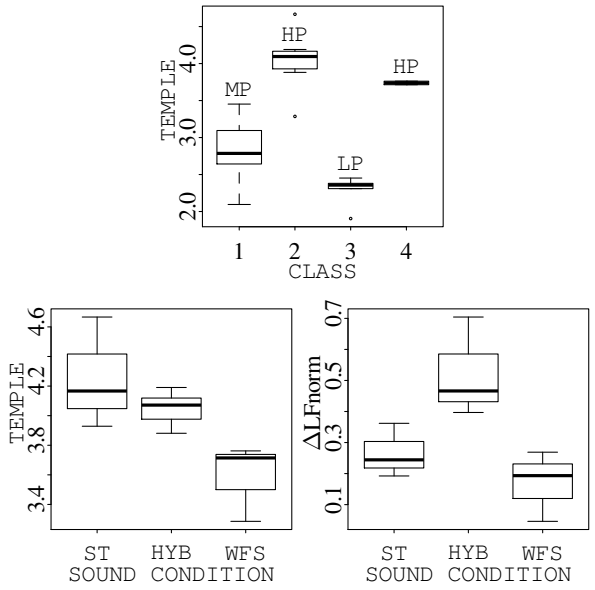

Figure 4: Boxplots of the `TEMPLE` score vs. the GMM classification `CLASS` (top), `TEMPLE` score vs. `SOUND CONDITION` for participants in group `HP` (bottom left), and the ΔLFnorm value vs. `SOUND CONDITION` for participants in group `HP` (bottom right).

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.