Proceedings of the 13th International Conference on Auditory Display, Montréal, Canada, June 26-29, 2007

SPATIAL AUDIO QUALITY EVALUATION: COMPARING TRANSAURAL, AMBISONICS AND STEREO

Catherine Guastavino[1], Véronique Larcher[2], Guillaume Catusseau[3] and Patrick Boussard[3]

[1] McGill University, CIRMMT and GSLIS
3459 McTavish, H3A 1Y1 Montreal, QC Canada
catherine.guastavino@mcgill.ca
[2] Sennheiser Research, 3239 El Camino Real Palo Alto, CA 94306, USA
LarcherV@sennheiser.com
[3] Genesis S. A., Domaine du Petit Arbois - B P 69
13545 Aix en Provence Cedex 4, France
patrick.boussard@genesis.fr

ABSTRACT

Two experiments were conducted to investigate perceptual differences between three sound recording and reproduction techniques, namely transaural, ambisonics and stereophony, in terms of spatial quality (Exp.1) and localization (Exp. 2) on a variety of sound material. Results indicate a strong contrast between ambisonics and the other two techniques. Specifically, ambisonics provides a good sense of immersion and envelopment but a poor localization and readability of the scene, while stereophony and transaural provide a precise localization and a good readability but lack immersion and envelopment. These results suggest that a trade-off between immersion and precision may be difficult to achieve using these techniques.

[Keywords: Multi-Channel Audio, Perceptual Evaluation]

1. INTRODUCTION

Sound quality evaluations for audio reproduction have traditionally been concerned with non-spatial attributes such as timbre or distortion while spatial attributes were extensively investigated in the context of room acoustics (see [3] for a review). However, the increasing use of multi-channel audio has recently motivated the study of spatial sound perception in the context of auditory displays to better understand how spatial attributes contribute to sound quality [2,3,5,6], ecological validity [4] and preference [1]. However, most studies focus on a specific recording or reproduction techniques. Our contribution is to compare three reproduction techniques in terms of ecological validity, spatial quality (Exp.1) and localization (Exp. 2).

Presented in this paper are the results of two listening tests in which transaural, ambisonics and stereophony were compared on a variety of source material. Double transaural is an extension of traditional transaural techniques ([8,9]) aiming at overcoming their limited sweet-spot and frequent front-back reversals [10]. To do so, frontal sources located in the front of the listener are rendered on a frontal stereo pair of speakers while sources located in the rear are rendered on an additional pair of speakers located behind the listener. Ambisonics and pairwise amplitude panning are documented in [11,12,7,13].

In Experiment 1, participants were presented with a reproduction of the same sound scene recorded using the three reproduction techniques and they were asked to evaluate the different versions of each recording using verbal descriptions and value scales. Experiment 1 investigates the influence of spatial presentation on listeners' perception of various attributes of the reproduced sound field. In Experiment 2, participants were presented with sounds positioned at different locations using double transaural, ambisonics and pairwise amplitude panning. Participants were asked to localize the sounds and rate the reproduction on value scales. Experiment 2 investigates the influence of spatial presentation on listeners' ability to localize sounds around them.

Both experiments resulted from a collaboration between Genesis (www.genesis.fr), the Laboratoire d'Acoustique Musicale (CNRS, Université Paris IV) and the Laboratoire de Mécanique et d'Acoustique (CNRS, Marseille).

2. EXPERIMENT 1: SPATIAL QUALITY EVALUATION

2.1. Methods

2.1.1. Reproduction techniques

Sound scenes were captured using three recording techniques simultaneously: binaural recordings were conducted using a Head Acoustics HS-II artificial head, first-order ambisonics recordings were conducted using a Soundfield ST 250 microphone, and plain stereo recordings were conducted using an ORTF setting (110 degrees angle and 17 cm between two cardioid microphones). The positioning of the above transducers was chosen so as to optimize their coincidence while minimizing occlusion, as shown in Fig. 1. The recordings were recorded on a 8-track Tascam DA-88 digital recorder, at a sampling rate of 48kHz.

Up to six loudspeakers were used for the playback. The stereo recordings were played back directly onto two loudspeakers located in front of the listener, at ± 30 degree azimuth. The binaural recordings were played back on the same loudspeakers, after transaural processing. The transaural decoder used was the default decoder delivered by Ircam with the Spat~ library, optimized for loudspeakers at ± 30 degree azimuth. Finally, the ambisonics recordings were decoded using Ircam Spat~ ambisonic decoder optimized for a playback on six loudspeakers - regularly spaced around the listener - including the two frontal loudspeakers mentioned above. The "in-phase" ambisonic decoder was selected as it is recommended for larger rooms and listening areas, preventing anti-phase signals to be fed to the loudspeaker opposite to the sound source.

The experiments took place in an anechoic chamber at the Laboratoire de Mécanique et d'Acoustique. The loudspeakers used were six Mackie HR824 studio monitors. They were equally spaced on a circle with a diameter of 4 m and hidden from view using acoustically transparent curtains, as shown in Fig. 2.

Figure 1. *Simultaneous recording of the 3 techniques: artificial head for transaural reproduction, Soundfield microphone for ambisonics reproduction and ORTF pair for stereophony.*

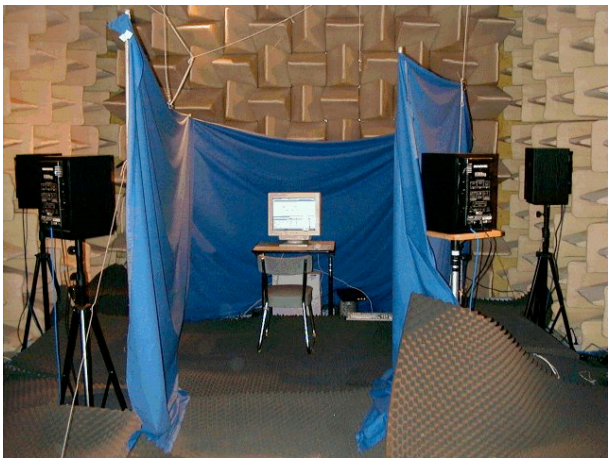


Figure 2. *Reproduction set-up where the six loudspeakers are hidden behind acoustically transparent curtains.*

2.1.2. Sound samples

Four auditory scenes were selected including an outdoor recording of traffic noise (30 seconds), and three indoor recordings, namely a car interior while driving (30 sec), people talking with background music at a reception (30 sec) and an excerpt of an electric guitar concert (10 sec).

2.1.3. Procedure

The graphical interface was programmed in jMax. On the first trial, participants were presented with a 30 sec loop of traffic noise recording. Instructions were given to direct their response strategy towards everyday listening situations, so that they would react, to some extent, as if they were in an actual situation i.e., in an ecological valid way [4]. A free verbalization task and a multiple comparison task were conducted: participants listened to the three reproduction methods as many times as desired, were asked to freely describe the three versions, choose which one(s) sounded the most similar to their everyday experiences, and justify their choice (see Appendix for full phrasing). This elicitation method, used in previous studies to investigate the sound quality of sound reproduction [3.4], was chosen to identify perceptually relevant features without constraining the answers into predefined categories. This open question addressed the ecological validity of the reproduction. It requires a strong familiarity with the sound material, and for this reason, it was only asked for the traffic noise recording.

On the following trials, participants were asked to rate the three reproduction methods (with three sliders on the computer screen corresponding to each reproduction method) for one the four sound samples along one of the 6 continuous scales listed in Table 1. The scales were constructed on the basis of previous research on spatial attributes [1,3,5,6]. The order of presentation was randomized within and across trials to nullify order effects. Completing the experiment took about an hour.

#	Scale	Phrasing	Range
1	Envelopment	The sonic environment	A little /

		sounds --- enveloping	very
2	Immersion	I feel --- immersed in the sonic environment	A little / very
3	Representation	Representation of the sonic environment	Poor / good
4	Readability	Readability of the scene	Poor / good
5	Realism	Naturalness, true to life	Not truthful / truthful
6	Overall quality	The quality of the reproduction is --	Poor / good

Table 1: Scales used in Experiment 1 (see Appendix 7.1.2 for original description in French language).

2.1.4. Participants and procedure

Eleven graduate students or staff from the Laboratoire de Mécanique et d'Acoustique and Genesis participated without pay in the experiment. They were aged between 25 and 50, studied or worked in the field of acoustics and can thus be considered as expert listeners.

2.2. Results

2.2.1. Qualitative analysis of the open question

Responses to the open question were classified into categories emerging for the spontaneous descriptions using the elicitation method presented in [3]. 43 phrasings were analyzed and grouped into semantic categories relating to Immersion/envelopment (8 occurrences), distance (6 occ.), rear sound (6 occ.), low frequencies (4 occ.), readability (4 occ.), "phasing effect" (4 occ.) and timbre (2 occ.). Semantic categories with fewer than 2 occurrences were excluded from the analysis. Ambisonics was described as very immersive (6 occ.), bassy (4 occ.), sounding close (3 occ.) with lots of rear sound (4 occ.). Transaural was described as immersive (2 occ.) and bright (1 occ.) but lacking rear sound (1 occ.) and sounding "inside the head" (1 occ.). A negative "phasing effect" related to instability to head movements was described (4 occ.) for transaural reproduction. Stereo was described as being frontal (3 occ.), sounding far (2 occ.), lacking rear sound (1 occ.) and muffled (1 occ.).
Regarding the selection task, transaural and ambisonics were selected 4 times each, while stereo was selected twice[1].

2.2.2. Statistical analysis of the ratings

A 3 (reproduction techniques) x 4 (sound samples) factorial ANOVA revealed a significant interaction effect of techniques·material ($F(6,780)=6.47$, $p<0.001$), as shown in Fig. 3. *Post-hoc* analyses were conducted using Tukey's HSD test. The only significant difference was observed between transaural and both ambisonics and stereo for the concert excerpt ($p=0.01$). A very significant effect of reproduction techniques was observed ($F(3,792)=10$, $p<0.0001$) and no significant effect of sound samples were observed ($F(3,792)=0.085$). Hence the results will be presented for all sound samples together.

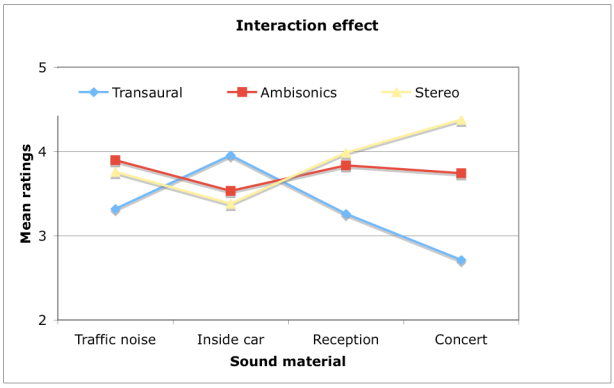


Figure 3. *Interaction effect of reproduction technique·sound samples. A significant difference was observed between transaural and both ambisonics and stereo for the concert excerpt (p=0.01).*

A one-way ANOVA was conducted to investigate effects of reproduction techniques for each of the 6 scales. The results are reported in Table 2. The ratings for each scale grouped by reproduction technique and averaged over all participants and sound samples are reported in Figure 4. Significant effects of reproduction techniques were observed for envelopment, immersion, readability, realism and global rating.

Post-hoc analyses were conducted using Tukey's HSD test. Ambisonics was rated as significantly more enveloping and more immersive than both transaural and stereo ($p=0.01$), but also significantly less readable than transaural and stereo ($p=0.05$). Regarding realism, stereo was rated as significantly more realistic than transaural ($p=0.001$). Regarding overall quality, stereo and ambisonics were rated significantly higher than transaural ($p=0.01$). No other significant differences were observed.

#	Scale	F(2,129)	p-value	Significance
1	Envelopment	7.22	0.001	Yes
2	Immersion	7.04	0.001	Yes
3	Representation	3.84	0.27	No
4	Readability	7.82	<0.001	Yes
5	Realism	5.58	0.004	Yes
6	Overall quality	14	<0.0001	Yes

Table 2: Results of the ANOVA comparing the 3 reproduction techniques (averaged over all participants and all sound sources) for each scale.

[1] One participant chose not to respond.

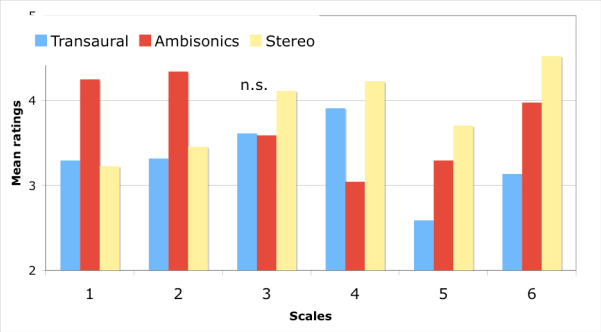


Figure 4. *Mean ratings for each scale grouped by reproduction technique. The ANOVAs revealed significant differences for all scales except 3 (see F and p values in Table 2).*

2.2.3. *Correlation between scales*

A moderate correlation was observed between Envelopment and Immersion (r=0.47, r^2=22%) suggesting moderate overlap between the two scales. All other correlation coefficients were below <0.4 suggesting that the scales measure different attributes.

3. EXPERIMENT 2: LOCALIZATION

3.1.1. *Reproduction techniques*

In order to investigate the accuracy of sound positioning for spatial recording techniques, a controlled and reproducible sound scene was created. It consisted of a monophonic sound playing on each of six loudspeakers regularly spaced around the sweet spot and placed in a typical conference room. The monophonic sound was not only recorded for the positions corresponding to the six loudspeakers (±30 degrees, ±90 degrees, ±150 degrees), but it was also recorded when reproduced at the position between two speakers using amplitude panning, bringing the number of characterized positions to twelve. Several monophonic sounds were recorded in that setting. This time, only binaural and ambisonics recordings were conducted. These recordings took place one at a time, thereby making it easier to position each microphone system at the same location. No plain stereo recording technique was investigated as none can efficiently capture positional cues of sources located far outside of their recording angle. Instead, pairwise amplitude panning was used as a reference for comparison.

The ambisonics recordings were played back using the same decoder as in Experiment 1. This time, the binaural recordings were decoded using a "custom" double transaural decoder, based on the decoder provided with Ircam Spat~ library. Our decoder was using the same transaural decoder as in Experiment 1, except that for sources located in the rear, the decoded channels were routed towards two loudspeakers located in the rear, in a symmetrical position to the loudspeakers used for the transaural reproduction of frontal sources. Therefore, up to four loudspeakers were used to play back the binaural recordings. It should be noted that artificial head recordings of complex sound scenes can generally not be decoded for double transaural reproduction since such a system would require segregating sources coming from the front from sources coming from the rear.

3.1.2. *Sound samples*

Four sound samples were selected to cover a wide range of spectrum and temporal evolution. All samples were 10 second long. They are described in Table 3 in terms of context and in the Appendix in terms of spectrum and waveform.

	Description
1	Synthetic white noise with slow amplitude modulation
2	Male spoken voice recorded in anechoic room
3	Synthetic bubbling sounds made of noise bursts
4	Musical phrase on a trombone recorded in anechoic room

Table 3: Description of the sound samples used in Exp. 2 (see Appendix 7.2. for more details).

3.1.3. *Participants and procedure*

The same set of 11 participants completed Experiment 2 in a separate experimental session separated by a week. Completing the experiment took about one hour and a half.
Sounds were positioned at the following angles: 0° (frontal source), 30°, 60°, 90°, 120°, 150°, 180°, 210°, 300°, 240°, 270°, 300°, 330° and 360°. Out of these twelve angles, only seven were tested for each participant to reduce the number of trials by excluding opposite angles (e.g. if using 30°, then 330°, i.e. -30°, was not tested and vice-versa). The order of presentation was randomized across trials to nullify order effects and counterbalanced across participants to cover all twelve angles. On each trial, participants were asked to localize the sound by selecting one of the twelve positions on a circle and then evaluate the ease of localization and the precision of the source on a continuous scale of 0 to 7.

3.2. Results

3.2.1. *Localization task*

The results of the localization test are presented in Figure 5 for each positioning technique. We computed the correlation between the actual reproduced angle and the perceived angle for each reproduction technique. Reported in Table 4 are the overall correlation coefficient and the coefficient of determination (r^2), which corresponds to the percentage of variance in perceived angle that is accounted by the variance in actual reproduced angle. Reported in Table 5 are the correlation coefficients for each sound sample.

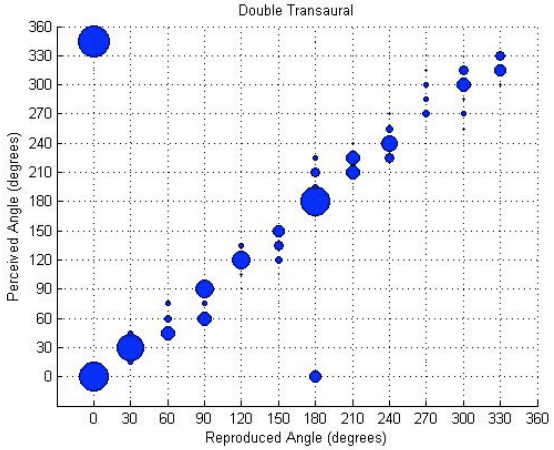
Technique	**Correlation coefficient**	**% of variance explained**	**Degree of correlation**
Double transaural	*r = 0.72*	*52%*	*Strong*
Ambisonics	r = 0.49	24%	Moderate

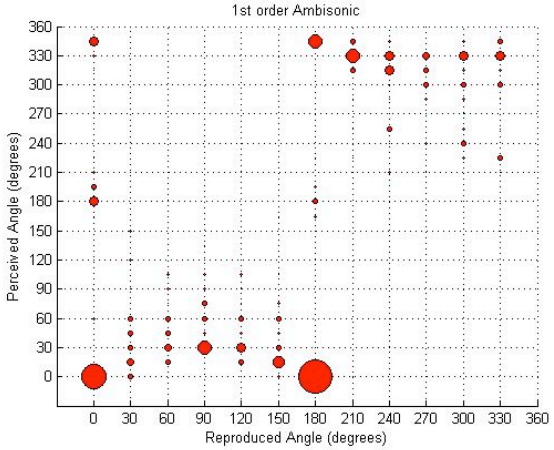
Pairwise Amplitude Panning	*r = 0.85*	*72%*	*Very strong*

Table 4: Correlation between the actual reproduced angle and the perceived angle for each reproduction technique (collapsed over all participants and all sound sources yielding 308 data points for each technique). Strong correlations are indicated in italics.

Technique	**White noise**	**Voice**	**Bubbles**	**Trombone**
Double transaural	*0.72*	*0.75*	*0.72*	*0.68*
Ambisonics	0.48	*0.63*	0.47	0.37
Pairwise Amplitude Panning	*0.88*	*0.80*	*0.86*	*0.87*

Table 5: Correlation between the actual reproduced angle and the perceived angle for each technique and for each sound source (77 data points for each technique). Strong correlations are indicated in italics.





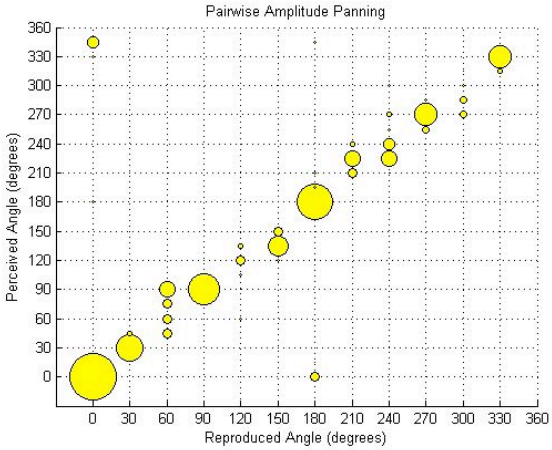


Figure 5. *Results of the localization test for transaural (blue), ambisonics (red) and pairwise amplitude panning (yellow) for all sound samples.*

As can be seen from the data shown in the tables 4 and 5 and in Figure 5, the accuracy of localization with ambisonics is overall significantly lower than for the pairwise amplitude panning and double transaural techniques. This is especially true for sound sources recorded on the sides. The rate of front-back confusions adds to this lower performance of ambisonics, since the rate is of 7% and 11% for pairwise amplitude panning and for the double transaural respectively, and reaches 38% for ambisonics. For the first two techniques, the confusions occur for sources reproduced directly in front or in the back of the listener. In the case of ambisonics, not only are the confusions for these positions more frequent, but confusions also occur for the neighboring positions of stimuli.

3.2.2. Ratings

One-way ANOVAs on the ratings for each scale, averaged over all participants and sound samples revealed a significant effect of reproduction technique on both the ease of localization ($F(2,921) = 86.3$, $p<0001$) and the precision of the source ($F(2,921) = 78.01$, $p<0.001$) as shown in Figure 6.

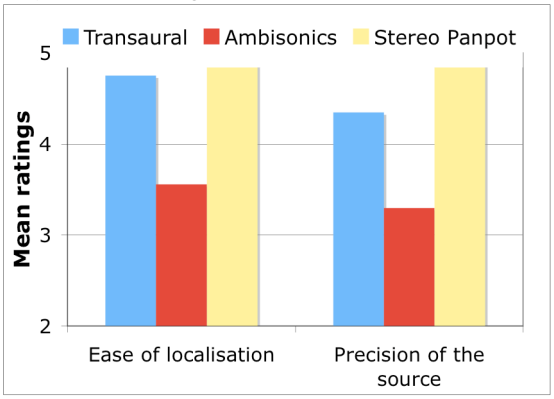


Figure 6. Mean ratings for each scale grouped by reproduction technique. The ANOVAs revealed a significant effect of reproduction technique.

Also observed a strong correlation between the ratings along the 2 scales ($r = 0.78$, $r^2 = 60\%$ of variance explained) suggesting redundancy across the two scales.

4. CONCLUSIONS

The main findings of Exp. 1 and 2 are summarized in Table 6. Results indicate a strong contrast between ambisonics and the other two techniques. Specifically, ambisonics provides a good sense of immersion and envelopment but a poor localization and readability of the scene, while stereophony and transaural provide a precise localization and a good readability but lack immersion and envelopment. These findings are in agreement with the analysis of binaural cues reported in [7] showing that binaural cues[1] for ambisonics are unstable compared to binaural cues for pair-wise (or triplet-wise) panning.

Reproduction technique	Strengths	Weaknesses
Transaural	Precise and easy localization Good readability	Poor realism and lack of immersion/envelopment
Ambisonics	Strong immersion and envelopment	Poor localization readability
Stereo / Panpot	Very precise localization	Lack of immersion/envelopment

Table 6: Characterization of the reproduction techniques.

On methodological grounds, results of Exp.1 suggest that the phrasing of the scale "representation" was too vague and did not help characterize the different reproduction techniques studied here.

Further analysis of the localization test will include comparing front-back confusion rates across techniques, and accuracy for sounds positioned between speakers as opposed to on the speakers. Directions for future research include investigating the spatial quality and localizability of Wave Field Synthesis, which may provide a good trade-off between immersion and precision.

5. ACKNOWLEDGMENTS

The writing of this paper was supported by FQRNT and CFI grants to Catherine Guastavino. The authors would like to thank Benoit Gauduin for resurrecting old jMax patches.

6. REFERENCES

[1] J. Berg & Rumsey, F. ''Spatial attribute identification and scaling by repertory grid technique and other methods,'' *Proceedings of the 16th AES International Conference on Spatial Sound Reproduction*, Audio Eng. Soc. 1999.

[2] S. Choisel & Wickelmaier, F. "Evaluation of multichannel reproduced sound: Scaling auditory attributes underlying listener preference," *J. Acoust. Soc. Am.*, vol. 121, no 1, pp. 388–400, 2007.

[3] C. Guastavino, & Katz, B. "Perceptual evaluation of multi-dimensional spatial audio reproduction," *J. Acoust. Soc. Am.*, vol 116, no 2, pp. 1105–1115, 2004.

[4] C. Guastavino, Katz, B., Polack, J-D., Levitin, D., & Dubois, D. "Ecological validity of soundscape reproduction," *Acta Acustica united with Acustica*, vol. 91, no. 2, pp. 333–341, 2005.

[5] F. Rumsey "Spatial quality evaluation for reproduced sound: terminology, meaning and a scene-based paradigm," *J. Audio Eng. Soc.* Vol. 50, no. 9, pp. 651–666, 2002.

[6] N. Zacharov & Koivuniemi, K. ''Audio descriptive analysis & mapping of spatial sound displays,'' *Proceedings of the 2001 International Conference on Auditory Display*, Espoo, Finland, July 29–August 1, 2001.

[7] V. Pulkki. "Evaluating Spatial Sound with Binaural Auditory Model," Proceedings of the International Computer Music Conference, pp. 73–76, Havana, Cuba, Sep. 2001.

[8] B. Gardner. "Transaural 3D Audio". *Technical Report 342*, M.I.T. Media Lab Perceptual Computing, July 1995.

[9] H. Moller. "Reproduction of artificial-head recordings through loudspeakers". *J. Audio Eng. Soc.* 37 (1/2):30-33, 1989.

[10] T. Funkhouser, Jot, J.-M. & Tsingos, N. "Sounds good to me! Computational Sound for Graphics, Virtual Reality, and Interactive Systems". *SIGGRAPH 2002 Course Notes*. http://www.cs.princeton.edu/~funk/course02.pdf

[11] J. Chowning, 1971. "The simulation of moving sources". *J. Audio Eng. Soc.* 19(1), 2-6.

[12] M. A. Gerzon, 1972. "Periphony: with-height sound reproduction". *J. Audio Eng. Soc.* 21(1), 2-10.

[13] J.-M. Jot, Larcher, V. & Pernaux, J-M. "A comparative study of 3-D ausio encoding and rendering techniques". *Proceedings of the 16th conference of the Audio Eng. Soc.*, pp. 281-300, Rovaniemi, 1999.

7. APPENDIX

7.1. Formulation of the questions in Exp. 1

7.1.1. Open question

"First you will be asked to listen to all three versions and select the one(s) that sounds the most like your everyday life experience. To do so, try to imagine that you are "there", in context. Closing you eyes might help. Please specify how you have made your choice?"

Original question in French:
Il s'agit tout d'abord de choisir parmi les 3 séquences qui vous sont présentées celle(s) qui vous semble(nt) la(les) plus proche(s) de votre expérience quotidienne. Pour cela, essayez de vous imaginer dans le lieu, de vous mettre en situation, éventuellement en fermant les yeux. Veuillez préciser pourquoi vous avez choisi cette (ces) sequence(s).

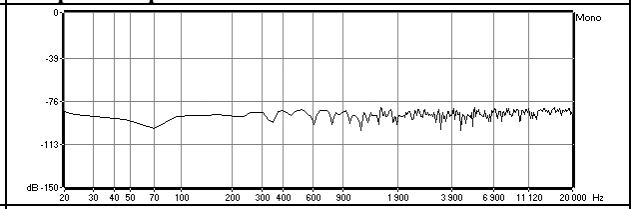
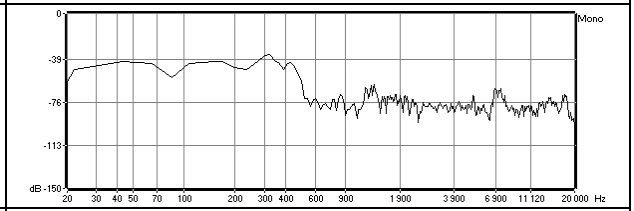
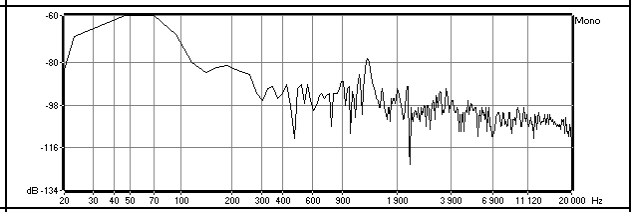
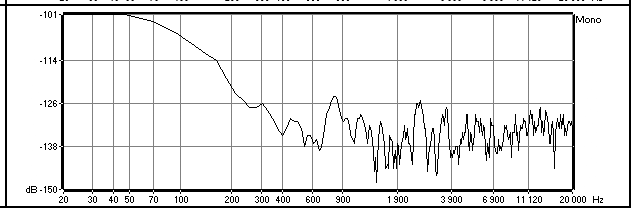
[1] The cues investigated in [7] are the ITDA (Interaural Time Difference Angle) and the TLDA (Interaural Level Difference Angle).

7.1.2. Scales

1. L'environnement sonore qui m'est présenté me semble : peu enveloppant / très enveloppant.
2. Je me sens : peu immergé / très immergé dans l'environnement sonore qui m'est présenté.
3. Je me représente l'environnement sonore : pas du tout / entièrement.
4. L'environnement sonore qui m'est présenté me semble : peu lisible / très lisible.
5. L'environnement sonore qui m'est présenté me semble : peu fidèle / très fidèle à une expérience réelle.
6. La restitution sonore me semble de qualité : très médiocre / très bonne.

7.2. Description of sound sample used in Exp. 2

	Description
1	Synthetic white noise with slow amplitude modulation
2	Male spoken voice recorded in anechoic room
3	Synthetic bubbling sounds made of noise bursts
4	Musical phrase on a trombone recorded in anechoic room

	Amplitude Spectrum
1	
2	
3	
4	

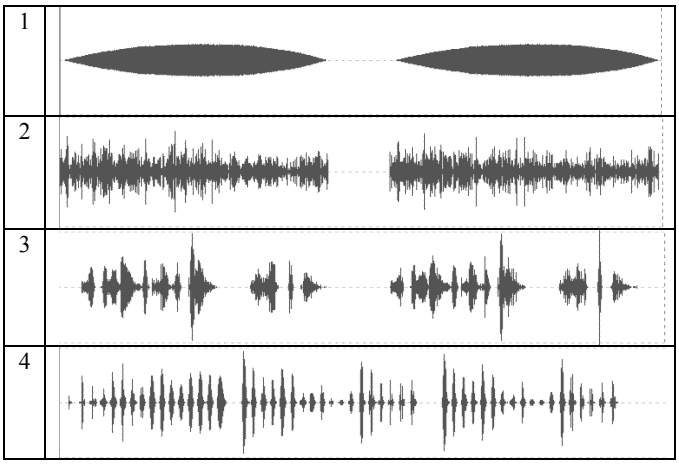
	Waveform
1	
2	
3	
4	

Table 7: Description of the sound samples used in Exp. 2 in terms of content, amplitude spectrum and waveform.