Should a movie have two different soundtracks for its stereoscopic and non-stereoscopic versions? A study on the front/rear balance

Etienne Hendrickx, Mathieu Paquier, Vincent Koehl

SHOULD A MOVIE HAVE TWO DIFFERENT SOUNDTRACKS FOR ITS STEREOSCOPIC AND NON-STEREOSCOPIC VERSIONS? A STUDY ON THE FRONT/REAR BALANCE

Etienne Hendrickx, Mathieu Paquier and Vincent Koehl

University of Brest, CNRS, Lab-STICC UMR 6285
6, avenue Victor Le Gorgeu, CS 93837
29238 Brest Cedex 3, France
etienne.hendrickx@univ-brest.fr

ABSTRACT

Few psychoacoustic studies have been made on the influence of stereoscopy on the sound mixing of movies. Yet very different opinions can be found among scientific, esthetical or technical communities. Some argue that sound needs to be mixed differently for stereoscopic movies, whereas others pretend that image has actually caught up with sound, that was already "three-dimensional" and should not therefore be affected by stereoscopy. In the present experiment, expert subjects were asked to achieve surround sound ambiance mixings for eleven short sequences presented in both stereoscopic and non-stereoscopic versions. The results suggest that the influence of stereoscopy on the front/rear balance strongly depends on the content of the sequence and only appears in a few specific situations.

Index Terms— Cinema, stereoscopy, sound mixing, balance, surround, 3D.

1. INTRODUCTION

1.1. Technical Considerations

In Hollywood, very different approaches can be found: for some sound engineers, stereoscopic images completely change our auditory perception and two different soundtracks are to be mixed for the 2D and 3D versions of a movie. Others argue that this influence is weak, even negligible.

Michael Semanick[1] affirms that there were many differences between the stereoscopic and the non-stereoscopic mixes of Tim Burton's "Alice in Wonderland": For the 3D version, they added some surround effects in areas and pulled music out into surround a bit more. They also pushed the reverberation and the surrounds of the backgrounds further, and panned a bit more the dialog [1].

---

[1] Two-time Academy Award winner sound mixer for "The Lord of the Rings: The Return of the King" (2003) and "King Kong" (2005).

Paul Martin Smith, editor of Eric Brevig's "Journey to the center of the earth", also defends the idea that the sound mixer should always work with the image projected in stereoscopy, because it influences the way one places sound sources in the space [2].

On the other hand, the mixing crew of Martin Scorsese's "Hugo Cabret" considers the 3D actual revival as a phenomenon concerning image rather than sound [3]. They argue that sound was already "in 3D with the surround speakers" and should not be too much affected by stereoscopy. James Cameron and its sound crew seem to share the same opinion. On "Avatar", they worked with the 2D version, and occasionally checked if their mixing worked with the stereoscopic version. According to them, they hardly made any modifications [4].

1.2. Scientific Considerations

The few studies that have been made so far on sound related to stereoscopic images have mainly focused on physical realism. They rely on the hypothesis that a sound reproduction closer to a real sound field (such as Wave Field Synthesis, Ambisonics, or binaural techniques) should increase even more the feeling of being part of the movie, in the same way that stereoscopic images increase the sense of presence of the audience, as shown by Ijsselsteijn *et al.* in [5]. However, it has not been proved yet that stereophonic images increased the sense of presence because it was closer to human perception, and besides several studies suggest that a physically realistic reproduction of sound may not be appropriate for movies: André *et al.* [6] played back a stereoscopic movie with different sound conditions, from stereo, that has a low spatial audiovisual coherence, to WFS, that has a very high spatial audiovisual coherence. Only 12 out of 33 participants felt that the sound condition had had an impact on their sense of presence, and they reported the WFS soundtrack as providing the lower sense of presence, when it was supposed to be the more physically realistic. However, only one sequence was used for the study (the first three scenes of an animation stereoscopic movie) with a sound reproduction system that was not 360° but only

frontal. Further psychological studies are therefore needed to confirm that a physically realistic reproduction of sound can truly increase or decrease the sense of presence [7]. Moreover, a recent study has shown that too much spatial information could reduce attention to narrated text [8]. Even if this study was carried out without images, it may be a clue to explain the results of André *et al.*.

1.3. Esthetical Considerations

For Chion [9], people are so accustomed to the conventions of cinema that they have become new references of reality. Rumsey [10] also hypothesizes that reproduced sound may have acquired its own standard of realism, different from natural listening.

The ventriloquism effect is a good example to support this idea: in theaters, the voices are most of the time reproduced on the central speaker. Yet the voices seem to match with the position of the actors, even when they are at the borders of the screen or off screen. Sometimes, breaking this "rule" and panning the dialog can even be perceived in a negative way [11].

However, It may sometimes be interesting to draw inspiration from reality and try to adapt some physical aspects of a real sound field to fit in a "traditional" reproduction system, like the 5.1 configuration. For example, Chion [9] suggests that surround gives more credit to ambiance sounds because it is sensorily more convincing, as it reproduces the fact that sound comes not only from front but also from behind.

Nevertheless, the goal of this study is not to determine whether a more physically realistic reproduction of sound can increase the sense of presence, but rather to study the differences of auditory perception that may occur when watching a movie in 2D or in 3D. The present study focuses on the influence that stereoscopy may have on the mixing of backgrounds. Subjects were asked to adjust the front/rear balance for eleven stereoscopic sequences, along with their non-stereoscopic version. If the results show that background surrounds are significantly pushed further when watching the sequences in 3D, it could mean that a stereoscopic image, as it is closer to human perception, motivates the subject to have a sound reproduction closer to reality as well, with no discrimination between the front and the rear sounds (as opposed to the "traditional" sound in 2D cinema, in which most of the energy comes from the front speakers [12]). Either way, it would suggest that new "conventions" (or "standards of realism") are needed, and that two different soundtracks have to be mixed for movies that are to be either projected in 2D or in 3D.

2. EXPERIMENTAL SETUP

2.1. Material

Eleven stereoscopic sequences were used for the test, along with their non-stereoscopic version (see Tab. 1).

Seven sequences were specially shot for the test, while the remaining four were taken from "Tonnerre de Brest", a 3D documentary by Pierre Souchar. All the sequences were shot with a Panasonic AG-3DP1 camera, and were chosen so that they would offer a variety of dynamics (static shots, dollying in, hand-held shots, etc.), types (close shots, medium shots, long shots, etc.) and categories (sea, city, crowd, interior, forest).

Also Guichardan had noticed that surrounds could be more or less efficient depending on the dramatic content of the stimulus used [13]. Though his experience only concerned 2D images, it was nonetheless decided to vary the "degree of dramaturgy" of our sequences, from sequences in which basically "nothing happens" (long shot of a quiet sea) to dialog or music scenes.

2.2. Recordings

It was also decided to use different recording and mixing techniques that are commonly used for professional shootings (see Tab.1).

3 sequences were recorded using Double-M/S (see Fig.1), one of the most established surround recording techniques for applications such as documentary sound and radio drama [14]. It consists of two cardioid microphones (one facing forward, and one facing backward) and one bi-directionnal microphone (angled 90°), whose signals are then matriced to generate multi-channel sound tracks.
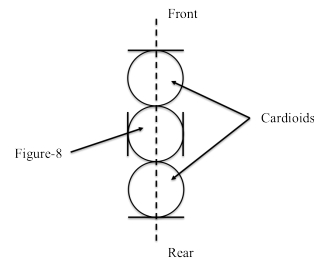
Fig. 1. Double-M/S system. The three microphones are supposed to be coincident.

4 sequences were mixed using decorrelated stereophonic backgrounds (one reproduced on the front speakers and the other reproduced on the rear speakers), which is a very common way of building ambiance sounds. Those backgrounds were recorded with an ORTF pair, which consists of two cardioid microphones spaced with 17 cm in a base angle of 110°, and which has been proved to yield a fine distance discrimination compared to other traditional arrays such as XY or M/S [15].

2 sequences were recorded using Double-ORTF (see Fig. 2), a system with four cardioids arranged in a perfect square shape, with an angle of 110 degrees between the two front microphones and between the two rear microphones.

	Content	Shot type	Recording System	Snapshot
1	Cellist playing Bach's third cello suite in a very reverberant church	Dollying in	Fukada Tree	
2	Cellist speaking, still in the same church	Close shot	Fukada Tree	
3	Interior of a childcare center, with backgrounds of children playing from outside	Long shot	Decorrelated Stereo	
4	Dialog in a café, with dialogs reproduced on the central speaker	Two shot	Decorrelated Stereo	
5	Same shot that 4) but with the dialogs pushed further in the central speaker (+ 6 dB)	Two shot	Decorrelated Stereo	
6	Two girls walking in a forest, shot from behind, with their footsteps reproduced on the central speaker	Hand-held close shot	Decorrelated Stereo	
7	Bridge with cars and a tram passing by	Long shot	Double ORTF	
8	Harbor with a boat in the foreground and several boats in the background	Long shot	Double-M/S	
9	Quiet sea, from the bow of a ship	Long shot	Double-M/S	
10	Bow of a ship, with two sailors maneuvering.	Long shot	Double-M/S	
11	Crowd in the street	Medium shot	Double ORTF	

Tab. 1. The eleven sequences, with their content, shot type, recording system and snapshot.

Many tests have pointed out the efficiency of this system for surround sound [16].
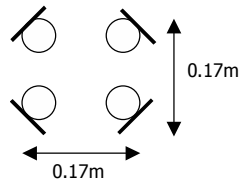
Fig. 2. Double-ORTF layout (from [17]).

All those recording setups were implemented using Schoeps CMC6 microphones with MK4 cardioid directivity capsules (except for the Double M/S system, which also included a Schoeps CMC6 microphone with a MK8 figure-eight directivity capsule), connected to a Sonosax SX-R4 audio recorder. A Neumann KMR 81 shotgun microphone was also used for the monophonic recording of dialogs and footsteps from sequences 4, 5, and 6.

At last, the two sequences in the church were recorded using a Fukada Tree. It is a non-coincident multichannel array that was the preferred system in the comparative study of Kassier, probably due to its pleasant spatial impression [18]. Several configurations exist for the Fukada Tree, and it was decided to use the same array as Hiekkanen *et al.* in [19] (see Fig. 3), which consists of three cardioid microphones forming a triangle for the front channels, and of two cardioids facing backward for the rear channels.
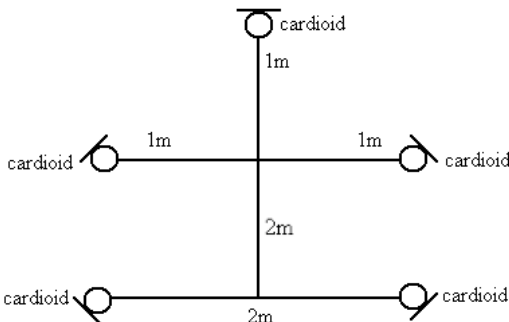
Fig. 3. Fukada Tree layout (from [19]).

The tree was implemented using five Schoeps CMC6 microphones with MK4 cardioid directivity capsules, connected to the inputs of a RME Fireface 800 interface.

2.3. Reproduction Setup

The impact of stereoscopy was studied on a "traditional" 5.1 mix. The listening test took place in the mixing auditorium of the Image & Sound department of the University of Brest. This room is especially designed for musical mixing and film postproduction. Five professional monitoring loudspeakers (PSI Audio 25-3) were arranged in 3/2 stereo configuration. The loudspeakers were fed by a RME Fireface 800 interface connected to a MacBook Pro computer, and the image was projected by an Epson EH-TW6000 projector, with Epson ELPGS01 3D active glasses.

All the sequences had been previously edited and mixed in that same auditorium by the experimenters. For each sequence, a front/rear balance was fixed by the experimenters, one for the stereoscopic version, and one for the non-stereoscopic version. The average of the two balances was then calculated to define for each sequence an original front/rear balance (nominally 0 dB) that would be retained as a reference for later analysis. This "0 dB-balance" did not necessarily mean having as much energy coming from the front speakers than coming from the rear speakers, which would have been irrelevant for some sequences such as sequences 1 and 2, in which there is only reverberation in the rear speakers.

The gains of the loudspeakers were adjusted so that they would individually produce at the listening position a sound pressure level of 85 dB, C-weighted, when playing a pink noise at − 20 dBFS RMS. This calibration is in accordance with the Dolby recommendations and was maintained during the whole process, from the original mixings to the final listening tests. For each sequence, a global level was set subjectively by the experimenters themselves, as it is often the case in subjective tests [20].

2.4. Subjects and protocol

The listeners were all paid volunteers from the Image & Sound course at the University of Brest. They were master's degree students, which means they already had a strong experience in critical listening and mixing, and could therefore be considered as "experts" [21].

The eleven stereoscopic sequences, along with their non-stereoscopic version, were presented to all the subjects. The order in which the 22 stimuli occured was random and different for each subject. They were asked to keep their 3D glasses at all time, even for the 2D sequences (during which the same image was sent to the left eye and to the right eye), to avoid an influence of the loss of brightness (that can go from 40 to 70% with active glasses). Each subject completed the test twice, with a fifteen-minute break between the two sessions. The average length of a session was about forty minutes. No subjects reported that they had experienced visual fatigue or discomfort due to prolonged 3D viewing, nor that the test had been too long or too demanding.
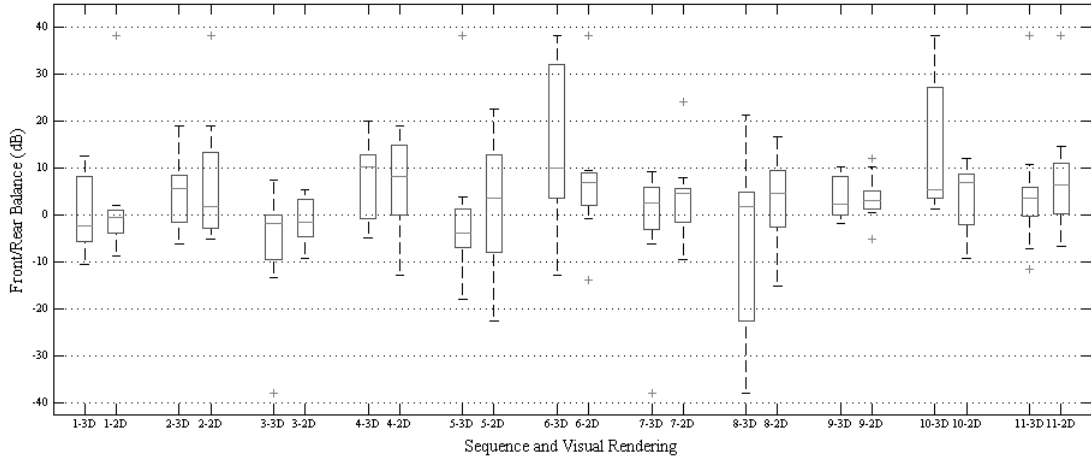
Fig. 4. Boxplots of the stereoscopic vs. non-stereoscopic balances of session 1, for the eleven sequences.
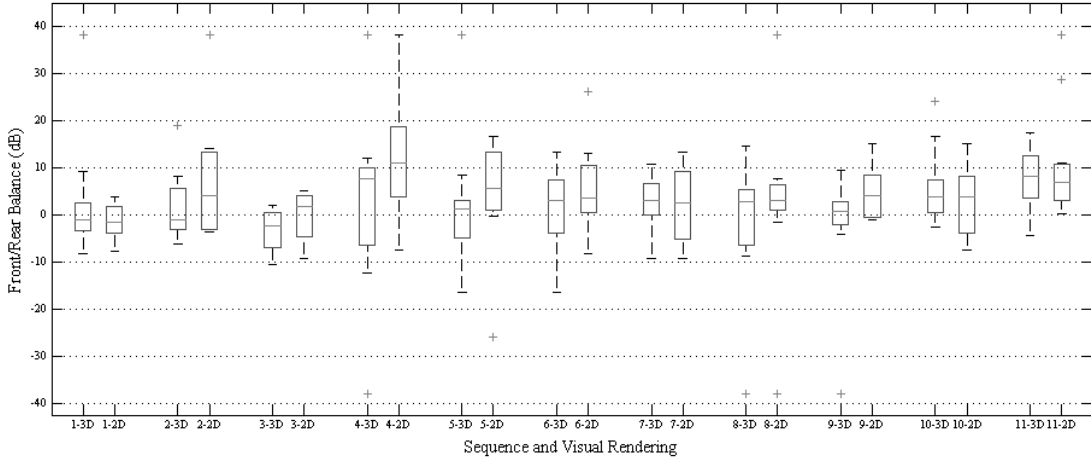
Fig. 5. Boxplots of the stereoscopic vs. non-stereoscopic balances of session 2, for the eleven sequences.

They were asked the question: "What front/rear balance would you like for this sequence?". They had to answer the question by setting themselves their proper balance using a knob from a Digidesign Command-8 controller, a MIDI interface that could provide 128 different steps for the balance. The knob was "infinite" and the gradation around it had been hidden, so that the listener did not have any visual or tactile feedback that could have influenced him. By turning the knob clockwise, the subject increased the level of front speakers while decreasing the level of surround speakers, until having only sound coming from front. A simple "sine-cosine" pan law was chosen for the evolution of intensity, which sounded more natural and closer to the subjects's everyday experience of mixing [22].

Let n be the MIDI value set by the subject, and G_F and G_R be respectively the amplification gain (expressed in dB) for the front and the rear speakers:

$$G_F = 20 \times \log_{10}\left[\sin\left(\frac{\pi}{2} \times \frac{n}{127}\right)\right] + 3\ dB$$

$$G_R = 20 \times \log_{10}\left[\cos\left(\frac{\pi}{2} \times \frac{n}{127}\right)\right] + 3\ dB$$

$$\Delta G = G_F - G_R = 20 \times \log_{10}\left[\tan\left(\frac{\pi}{2} \times \frac{n}{127}\right)\right]$$

For example, if the subject set the knob at its middle course ($n = 64$), then $G_F = G_R \approx 0\ dB$. Neither the level of the front speakers nor the level of the rear speakers were modified, which meant the subject chose the balance as it was initially mixed by the experimenters. If the subjects turned the knob to its maximal position ($n = 127$), then $G_F = +3\ dB$ and $G_R = -\infty$, with sound only coming from front. However, to avoid dealing with infinite numbers and

forbidden values, the subject could only set n between 1 and 126.

Each sequence was about 30 seconds long and automatically looped. When the subject was satisfied with his balance, he had to press a "push" button to go to the next sequence. Each sequence was initially presented with a random front/rear balance. Playback and data capture from the knob were controlled by a software implemented in Max/MSP.

For sequences 4, 5, and 6, the subjects could not modify the level of the dialogs and the footsteps that were reproduced on the central speaker.

3. RESULTS

Statistical analysis was performed, using ΔG as variable, to determine if there were any significant differences of balance between "stereoscopic mix" and "non-stereoscopic mix". As the distributions were not normal, a non-parametric statistical test appropriate to repeated measures had to be used: the Wilcoxon signed-rank test [23].

When analyzing the first session of the test (see Fig. 4), only the tenth extract (long shot of the bow of a ship, with two sailors maneuvering) was shown as having a significant difference between its stereoscopic and non-stereoscopic mixes (p = 0,0322), with medians equal to + 5.4 dB for the stereoscopic mix and + 7.0 dB for the non-stereoscopic mix. However, the difference was not significant anymore (p = 0.0645) if one did not take into account the results of subject 3 (+ 38.2 dB for the stereoscopic mix, which corresponds to the maximal value the knob can take, and + 8.9 dB for the non-stereoscopic mix). Besides, the results of subject 3 were far less extreme in the second session (+ 4.9 dB for the stereoscopic mix and − 7.6 dB for the non-stereoscopic mix).

Once the first session was finished, the subject would take a fifteen-minute break and would then complete the test a second time (with the sequences in a different order). The analysis of the second session showed significant differences for three sequences (see Fig. 5), with the non-stereoscopic mixes always more frontal than the stereoscopic mixes. Sequence 2 (close shot on a cellist speaking in a church) was significantly different (p = 0,0098), with medians equal to − 1.2 dB for the 3D version and + 4.1 dB for the 2D version. Sequence 4 (two shot of a dialog in a café, with dialogs reproduced on the central speaker) was significantly different (p = 0,0488), with medians equal to + 7.6 dB for the 3D version and + 11.1 dB for the 2D version. At last, sequence 9 (long shot of a quiet sea, from the bow of a ship) was significantly different (p = 0,0186), with medians equal to + 0.8 dB for the 3D version and + 4.1 dB for the 2D version. All the other sequences were not significantly different, including sequence 10 (p = 0,1309).

4. DISCUSSION

When there are significant differences, the 2D mix is always more frontal than the 3D mix, as supported by some sound engineers (see section 1.1.). However, it only concerns a few sequences: three significant differences for the second session, and one questionable difference for the first session.

Many subjects reported that they had trouble mixing complex sequences, such as sequence 7 (bridge with a lot of cars and a tram passing by) or sequence 11 (large crowd in a street). Throughout the two sessions, they would always notice new auditory or visual objects and would try to take them into account in their balance. Sequence 1 (the tracking shot in the church) was also problematic for the subjects, as they had trouble setting a balance that worked for the entire sequence. Most of them would have preferred to mix it dynamically, having more and more sound coming from front as the shot gets closer to the cellist. On the other hand, the three significant sequences of session 2 were very simple scenes: they were static shots, with few sound or visual cues, and very limited movements.

More significant differences were found in the second session than in the first one. It suggests that the subjects may have to go through a phase of learning in order to give relevant results. Besides, many subjects reported that it took them several sequences, sometimes an entire session, to realize that some sequences were actually in 2D (they were probably fooled by the 3D glasses that they were wearing permanently). More sessions are therefore needed to assert the results of this study.

5. CONCLUSION

The results suggest that the influence of stereoscopy on the front/rear balance strongly depends on the content of the sequence and only appears in a few specific situations. This is in accordance with the statements of the sound engineers of "Avatar", who claimed they had hardly made any modifications for the stereoscopic mix [4]. Stereoscopy tended to have an influence for simple sequences. For complex scenes, the number of parameters taken into account by the subjects for their balances became too important and significantly reduced the influence of the visual rendering.

6. ACKNOWLEDGMENTS

7. REFERENCES

[1] V. Gambier. Nouvelle approche pour la bande sonore d'un film en relief. *Graduate Research Project, Louis Lumière College*, page 55, 2010.

[2] B. Krohn. Entretien avec Paul Martin Smith. *Les Cahiers du Cinéma*, vol. 647, pages 16-19, 2009.

[3] M. Coleman. The sound of Hugo. http://soundworkscollection.com/videos/hugo.

[4] M. Coleman. The sound of Avatar. http://soundworkscollection.com/videos/avatar.

[5] 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, p. 298-311, 2001.

[6] C. R. André, M. Rébillat, and B. F. G. Katz. Sound for 3D cinema and the sense of presence. *Proceedings of the 18th International Conference on Auditory Display*, 2012.

[7] C. André, J. Embrechts, and J. G. Verly. Adding 3D sound to 3D cinema: Identification and evaluation of different reproduction techniques. *Audio Language and Image Processing*, pages 130-137, November 2010.

[8] M. Schmidt, S. Schwartz, and J. Larsen. Interactive 3D audio: enhancing awareness of details in immersive soundscapes? *133rd AES Convention*, Preprint 8780, October 2012.

[9] Michel Chion. L'audiovision. Armand Colin, pages 93-128, 2005.

[10] F. Rumsey. Spatial Quality Evaluation for Reproduced Sound : Terminology, Meaning, and a Scene-Based Paradigm. *J. Audio Eng. Soc.*, Vol. 50, No. 9, 2002.

[11] Michel Chion. Un art sonore, le cinema. Les Cahiers du Cinéma, page 122, 2003.

[12] L. Jullier. Le son au cinéma. Les Cahiers du Cinéma, page 17, 2006.

[13] T. Guichardan. Interactions between surround level and the immersive feeling in the multichannel movie experience. *111th AES Convention*, Preprint 5454, 2001.

[14] H. Wittek, C. Haut, D. Keinath. Double M/S – a surround technique put to test. *Tonmeistertagung*, pages 1-3, 2006.

[15] C. Hugonnet, and J. Jouhaneau. Comparative spatial transfer function of six different stereophonic systems. *82nd Audio Eng. Soc. Conv.*, London, Preprint 2465(H-5), 1987.

[16] A. Kornacki, B. Kostek, P. Odya, and A. Czyżewski. Problems Related to Surround Sound Production. *110th AES Convention*, Preprint 5374, 2001.

[17] A. Czyżewski, A. Kornacki, and P. Odya. Some rules and methods for Creation of Surround Sound. *21st AES Conference*, 2002.

[18] R. Kassier, H. Lee, T. Brookes, and F. Rumsey. An informal comparison between surround-sound microphone techniques. *118th AES Convention,* Preprint 6429, 2005.

[19] T. Hiekkanen, T. Lempiäinen, M. Mattila, V. Veijanen, and V. Pulkki. Reproduction of virtual reality with multichannel microphone techniques. *122nd AES Convention*, Preprint 7070, 2007.

[20] AES recommended practice for professional audio – subjective evaluation of loudspeakers. *J. Audio Eng. Soc.*, 1996 (reaffirmed 2007).

[21] ISO 8586-2, Sensory analysis – General guidance for the selection, training and monitoring of assessors – Part 2: Experts. *International Organization for Standardization*, (2008).

[22] D. Griesinger. Stereo and Surround Panning in Practice. *112th AES Convention*, Preprint 5564, 2002.

[23] M. Hollander, and D. A. Wolfe. Nonparametric Statistical Methods. Wiley-Interscience, 1999.