Qualitative evaluation of Wave Field Synthesis with expert listeners

A. Pras[1], E. Corteel[2], C. Guastavino[1]

[1] *McGill University and CIRMMT (Centre for Interdisciplinary Research on Music Media and Technology), Email: amandine.pras@mail.mcgill.ca; catherine.guastavino@mcgill.ca*

[2] *sonic emotion, 42 bis rue de Lourmel, Paris, France, Email: etienne.corteel@sonicemotion.com*

Introduction

Wave Field Synthesis (WFS) is a spatial sound rendering technique that enables the reproduction of target sound fields over an extended listening area [1]. WFS relies on the Kirchhoff-Helmholtz integral which states that any sound field can be reproduced in a subspace using a so-called continuous distribution of secondary sources located at the boundaries of said subspace. Practical implementations typically use a finite number of regularly spaced speakers located in a line. The underlying simplifications result in physical inaccuracies in the reproduced sound field. The subjective impact of these physical inaccuracies has been explored in previous studies mostly in terms of source localization [2][3][4]. It has been shown that a loudspeaker spacing of 15 to 20 cm is usually sufficient to ensure localization accuracy and to limit localization blur in an extended listening area. However, these studies mostly considered the case of a unique static sound source and static listeners.

Loudspeaker spacing is a key factor in the design of Wave Field Synthesis rendering systems that determines the required number of loudspeakers for a given listening area and room size. The finite loudspeaker spacing in practical Wave Field Synthesis sound reproduction systems creates spatial aliasing artifacts above the so-called aliasing frequency. These artifacts can be described as spatial overlapping components of the target sound field that may result in sound color variation artifacts for moving listeners. The use of partial decorrelation of loudspeakers' alimentation signal to reduce the amount of sound color variation has recently been proposed [5]. This partial decorrelation of loudspeakers' alimentation signal is referred to as diffusion throughout the paper.

This papers reports the result of a study conducted with 5 expert listeners who were presented with complex auditory scenes reproduced over 8 or 24 loudspeakers, with or without diffusion at high frequencies. They were asked to compare the 4 different reproductions and freely describe the perceived difference in their own words. Traditionally, quantifying perceptual attributes involves rigorous subject training to minimize differences among subjects and to identify small differences between parameterized stimuli. However, in the absence of documented subjective dimensions for WFS, a free exploratory approach was considered more appropriate to allow participants to define their own attributes rather than impose pre-defined factors of interest. An experimental protocol was designed to elicit relevant features by analyzing spontaneous verbal descriptions without constraining the answers into categories pre-defined by the experimenter.

Description of the system

Figure 1: **WFS playback system**

The WFS system used for the experiment is displayed in Figure 1. 24 closely spaced (13 cm) loudspeakers are installed above a projection screen at a height of approximately 2.3 m. The loudspeakers were equalized to exhibit a flat frequency response between 100 Hz and 20 kHz. Two subwoofers installed on each side of the screen were used at lower frequencies.

The WFS rendering could be modified in real time to switch between the 4 following rendering methods:

- 24 loudspeakers without diffusion (conventional WFS)
- 24 loudspeakers with diffusion
- 8 loudspeakers without diffusion
- 8 loudspeakers with diffusion

In the 8 loudspeaker setting, only 1 out of every 3 loudspeakers receives input signals, corresponding to a loudspeaker spacing of 39 cm. The diffusion filter was only applied above the aliasing frequency (600 Hz for 8 loudspeakers, 1800 Hz for 24 loudspeakers) so that it represents 50% of the energy at high frequencies. Each output receives an independently generated diffusion filter in

order to introduce partial decorrelation of the output channels.

All reproduction methods were configured such that the rendering level is the same in the listening area for all virtual sources.

Methods

Five expert listeners, aged 23 to 31, participated in the experiment, namely a movie production mixer, a music production sound engineer, 2 researchers in spatial audio and a musician. The participants served without pay.

For the purpose of this study, we edited and mixed 5 auditory scenes, consisting of 1 to 3 different sources (details presented in Table 1). These sound scenes were 1 to 2 minutes long, but most participants focused on shorter excerpts for the comparison. We used sound files from the LucasFilm library and personal recordings of isolated sources. To minimize experimental bias, we didn't use any sound effects (equalization, dynamic compression or artificial reverberation) during the mixing process.

Type	# sources	Back ground	Fore ground	Moving sources	Distant source
Outdoor	1	Waterfall			
Indoor	2		Speech		Percuss ions
Indoor	2	Guitars	Singing voice		
Outdoor	2	Forest Birds	Flute	Steps	
Outdoor	3	Kids playing		Steps Car	Drums

Table 1: Details of the sound sources for each auditory scene.

Procedure

The participants were presented with a reproduction of the same sound scene over four different systems in a randomized order: a) 24 loudspeakers without diffusion, b) 24 loudspeakers with diffusion, c) 8 loudspeakers without diffusion and d) 8 loudspeakers with diffusion. The participants were asked to freely describe the four versions. They could listen to the four reproduction systems as many times as desired and they were free to move in the listening space. However, participants found it difficult to listen while moving when the sources were also moving, and thus remained in a given position for these auditory scenes. Most participants compared the 4 versions pairwise, on two different excerpts of each scene (selected by the participants), resulting in 60 pairwise comparisons on average per participant. The listening test lasted for 2 hours.

Analysis of the verbal data

We conducted a content analysis of the spontaneous descriptions of the 5 sound scenes over the 4 reproduction systems using the constant comparison method [6]; see also [7] for spatial audio quality evaluation. Given the small number of participants and scenes, the number of phrasings for each scene or scene type was too low to be presented in isolation. Hence the results are collapsed over all auditory scenes.

A total of 114 phrasings were classified into 10 semantic categories emerging from the data. 74 of these phrasings referred to the effect of diffusion and 43 referred to the effect created by the number of loudspeakers. Synonyms were grouped together, as well as linguistic devices constructed on the same stem e.g., ''bright,'' ''brightness''. Lexical devices belonging to the same semantic field were grouped into semantic themes. Two coders independently combined semantic themes into larger semantic categories relating to Image width (28 occ. including source width for single-source scene and image width in the presence of multiple sources), Source distance (23 occ.), Stability to movement (18 occ.), Coloration (14 occ.), Phasing (9 occ. including phase problems, flange effect), Image depth (5 occ.), Image precision (5 occ.), Natural feel (4 occ.), Elevation (4 occ.) and Resonance (4 occ.). Finally, all occurrences in each category were counted. For each category, the distribution of phrasing is represented in Figure 2 for the effect of diffusion and the effect of the number of loudspeakers used for playback.

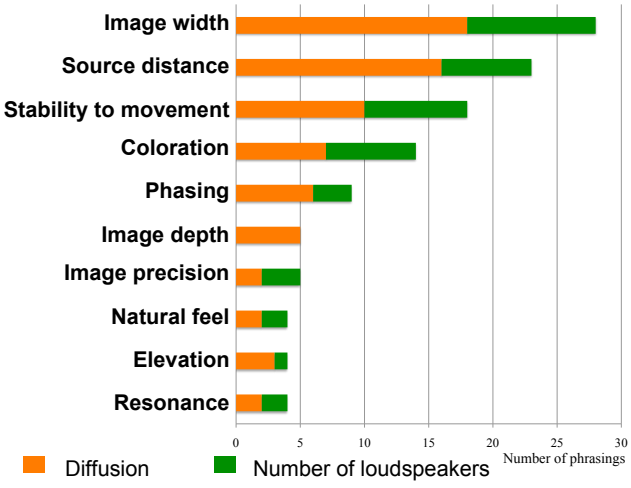

Figure 2: **Distribution of the verbal descriptions by semantic scales grouped as a function of varying the number of loudspeakers (in green) or diffusion (in orange).**

Presented in Table 2 are the opposite terms used by participants to describe each semantic scale. For some scales, only one term was used and so only one is reported here.

Semantic scale	**Opposite terms**	
Image width	Focus	Large
Distance	Close	Distant
Stability to movement	Unstable	Stable
Coloration	Low	High

Phasing		Phasing
Image depth	Flat	Deep
Image precision	Imprecise	Precise
Natural feel	Unnatural	Natural
Elevation		Elevation
Resonance		Resonance

Table 2: details of the opposite terms used to describe the semantic scales.

Presented in Figure 3 are the numbers of occurrences of spontaneous descriptions for the reproduction methods grouped by diffusion (with or without) and collapsed over the number of speakers, within each category. Opposing terms are represented on opposite sides of the graph. It can be seen that the image width, the source distance and the elevation increase with diffusion. As well, the sources become sensitive to movement with diffusion. To a lesser extent, diffusion adds some phasing and coloration in higher frequencies, increases the image depth and makes the sound scene unnatural and imprecise. Furthermore, a resonance artifact has been mentioned twice without diffusion.

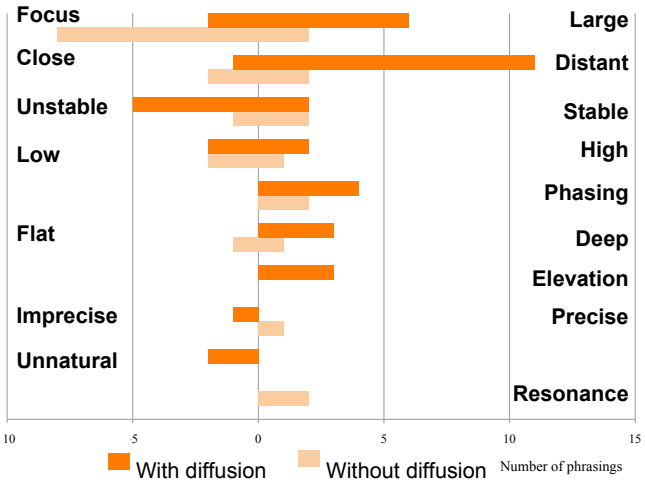

Figure 3: Distribution of occurrences of spontaneous descriptions for the reproduction methods grouped by diffusion (with or without).

Presented in Figure 4 are the numbers of occurrences of spontaneous descriptions for the reproduction methods grouped by number of loudspeakers (24 or 8), collapsed over diffusion conditions, within each category. Opposing terms are represented on opposite sides of the graph. Discriminating categories include Distance (sources are perceived as further with 8 loudspeakers) and Stability (sources are less stable on 8 compared than on 24). Furthermore, presentation on 8 loudspeakers resulted in higher coloration and resonance artifacts. To a lesser extent, the image was described as imprecise with 8 loudspeakers, as opposed to natural and precise with 24 loudspeakers. An elevation effect was mentioned once.

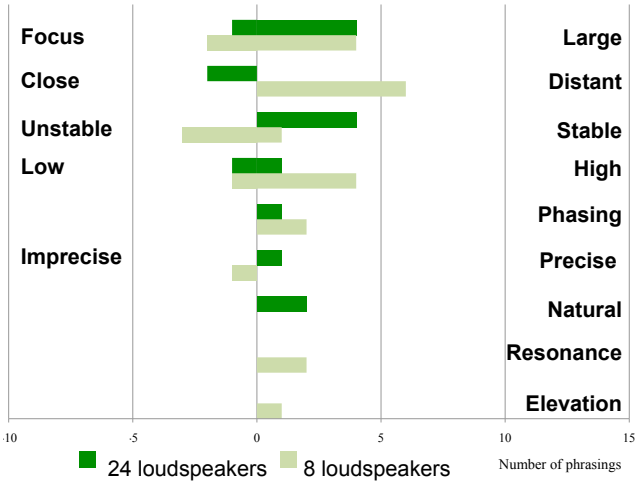

Figure 4: Distribution of occurrences of spontaneous descriptions for the reproduction methods grouped by number of speakers (8 or 24).

Discussion and further work

The verbal data analysis reveals a strong effect of diffusion. Diffusion has been found to enhance distance perception and less significantly image depth. This is consistent with results of Corteel et al. in [5]. In their experiment, an ABX protocol was used to evaluate the discrimination between electrodynamic loudspeakers or Multi-Actuator Panels with or without diffusion. Subjects reported that distance was the most important perceptual dimension for discrimination. This can be regarded as a positive impact of diffusion since it may limit localization of auditory events directly on the loudspeakers.

Diffusion also increases image width and reduces image stability. This can be regarded as an undesirable artifact. However, image stability was not significantly affected with diffusion. Additionally, diffusion affects coloration and may emphasize phasing artifacts. However, phasing artifacts were mentioned for the four different systems. This aspect needs to be further investigated.

Diffusion filter shape, frequency range and level should thus be adapted to limit artifacts while benefiting from distance and depth enhancement.

The number of loudspeakers was found to affect distance, stability and spectral artifacts. It should be noted however that changes in distance perception have only been mentioned when comparing 8 loudspeakers with diffusion versus 24 loudspeakers with diffusion. Thus, the reduction of the number of loudspeakers emphasizes the effect due to the diffusion regarding source distance. The results on spectral artifacts are consistent with Wittek et al. [8] who reported that the perceived coloration increases with increasing loudspeaker distances.

Interestingly, the descriptors derived from this free exploratory approach yielded semantic scales that were very similar to those in other spatial sound reproduction studies, with a different methodology and different reproduction

techniques [7] [9] namely distance, naturalness, depth, stability, and image precision. It is encouraging to note that a certain consensus begins to emerge in the field of spatial sound reproduction for perceptual attributes relating to spatial features, although the semantics of these terms vary across languages and may give rise to different interpretations (for a review of terminology and meanings of spatial attributes, cf. Rumsey, 2002 [11]).

Further research will use the semantic scales derived from this qualitative evaluation to systematically investigate the effect of diffusion and number of loudspeakers on a larger number of shorter excerpts with more participants. The excerpts will be identified on the basis of the selection used by our experts in this study. By systematically varying the complexity of the auditory scenes (number of sources, fixed or moving sources, depth), we will be able to characterize the optimal setting for each type of application as a function of the complexity of auditory scenes and properties of the sound sources.

Acknowledgments

This research was supported by FQRNT and CFI grants to Catherine Guastavino, namely FQRNT (NC-113581) and CFI (LOF-11367). The authors would also like to thank Aaron Rosenblum for his comments on an earlier draft.

References

[1] Berkhout A. J., de Vries D. and Vogel P., "Acoustic Control By Wave Field Synthesis", Journal of Acoustical Society of America, vol. 93, pp 2764-2778, 1993.

[2] E. N. G. Verheijen. "Sound Reproduction by Wave Field Synthesis". PhD thesis, TU Delft, Delft, Pays Bas, 1997.

[3] de Bruijn W., "Application of Wave Field Synthesis in Videoconferencing". PhD thesis, TU Delft, Delft, Pays Bas, 2004.

[4] Sanson J., Corteel E., Warusfel O., "Objective and subjective analysis of localization accuracy in Wave Field Synthesis", 124th Convention of the Audio Engineering Society, Amsterdam, Netherland, 2008.

[5] Corteel E., N' Guyen K-V., Warusfel O., Caulkins T., Pellegrini R., "Objective and subjective comparison of electrodynamic and MAP loudspeakers for Wave Field Synthesis", AES 30th international conference, Saariselkä, Finland, 2007.

[6] Glaser, B. G. The Discovery of Grounded Theory: Strategies for Qualitative Research, Chicago, USA, 1967.

[7] Guastavino, C., & Katz, B. "Perceptual evaluation of multi-dimensional spatial audio reproduction", Journal of the Acoustical Society of America, vol. 116(2), pp 1105-1115, 2004.

[8] Wittek, H., Rumsey, F, Theile, G., "On the sound colour properties of wavefield synthesis and stereo", 12[3rd] Convention of the Audio Engineering Society, New York, NY, USA, 2007.

[9] Berg, J., and 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 Engineering Society, 1999.

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

[11] Rumsey, F. ''Spatial quality evaluation for reproduced sound: terminology, meaning and a scene-based paradigm,'' J. Audio Eng. Soc. 50, pages 651 – 666, 2002.

[12] Bech, S. Perceptual Audio Evaluation Theory, Method and Application, Chichester, England: John Wiley & Sons, 2006.