Audio Engineering Society

Convention e-Brief 652

Presented at the 151st Convention
2021 October, Online

This Engineering Brief was selected on the basis of a submitted synopsis. The author is solely responsible for its presentation, and the AES takes no responsibility for the contents. All rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Audio Engineering Society.

Perceptual evaluation of a new, portable three-dimensional recording technique: "W-Ambisonics"

Xuan Lu[1], Sungyoung Kim[1], Miriam Kolar[2], and Doyuen Ko[3]

[1]*ECET, Rochester Institute of Technology, Rochester, NY, 14623, USA*
[2]*Department of Music, Amherst College, USA*
[3]*Audio Engineer Technology, The Mike Curb College of Entertainment and Music Business, Belmont University, USA*

Correspondence should be addressed to Xuan Lu (liebelux@gmail.com)

ABSTRACT

In order to exploit strengths and avoid weaknesses of the First Order Ambisonics (FOA) microphone technique, we devised a new, portable 3D microphone recording technique, "W-Ambisonics." This new technique incorporates a spaced stereo cardioid microphone pair (for frontal information) with two FOA microphone arrays (for lateral, rear, and height information). In W-Ambisonics, two FOA microphones are spaced 17 cm apart to capture and represent interaural cues precisely, with two cardioid microphones spaced 50 cm apart, 50 cm in front, which improves frontal directionality. Combining these two microphone pairs enables the translation of recorded audio into various reproduction formats according to practical limitations in reproduction peripherals. The design focus of this technique was efficiency in the recording stage and scalability in the reproduction stage. We conducted three perceptual experiments whose results show that the W-Ambisonics method enables improved lateral localization, provides comparable sound quality to the conventional spaced array technique, and translates spacious yet precise sound images in listening evaluations of a binauralized headphone rendering. The W-Ambisonics microphone technique is practical, precise, and scalable across multiple reproduction scenarios, from binaural to multichannel systems.

1 Introduction

Ambisonics is a spatial audio technique which aims to analyze, synthesize and reconstruct physical behaviors of a sound field. This technique has been researched for decades since the pioneering work by Gerzon [1]. This technique can be understood as a system of full periphonic directional sound pickup, storage, processing and reproduction around an origin position of a sound field based on the theory of spherical harmonics [2, 3]. The technique can be distinguished by the order of spherical harmonics in use. For example, a First Order Ambisonics (FOA) system encodes and decodes a sound field using the 0^{th} and first order spherical harmonics. Higher Order Ambisonics (HOA) extends the range of spherical harmonics.

The characteristics of an Ambisonics system can be summarized in terms of the four following features:

- It is a single-point recording technique. An FOA microphone array captures the entire sound field by using four coincidentally positioned microphone capsules. Each capsule corresponds to both

sound pressure and the three orthogonal components of velocity at the center of the sound field respectively. The captured signals represent the 0^{th} and first spherical harmonics [2].

- In contrast to object-based methods, it is independent of sound source number or speaker configuration, and can be scaled to any desired reproduction system without requiring specific up- or down-mix [4].
- It can provide a relatively stable phantom sound image and good immersive experience [4].
- As a single-point recording method, it is much smaller, more portable, and requires less setup time than large microphone arrays, which are typically more costly [5].

Despite the advantages of Ambisonics microphones, previous studies [6, 7, 5, 8] of 2D and 3D microphone techniques comparison show that Ambisonics-based recordings typically received lower perceptual ratings than spaced microphone techniques (e.g., the 3D extension of the Hamasaki Square), or near-coincident microphone techniques (e.g., ORTF-3D), specifically in localization precision and timbre coloration. Moreover, Ambisonics techniques, particularly FOA, present much wider sound images of *direct sound sources*, resulting in a 180-degree “wrap-aroud” of frontal sound, which contrasts with customary “cinematic” perspectives in stereo audio. This widened presence of frontal images can be perceived as “unnatural” [5, 9].

In order to improve on these FOA limitations—and to better utilize the Ambisonics advantage for ambient sound and reverberation—we developed a new recording technique, “W-Ambisonics,” that we propose here. The method combines a pair of FOA microphone arrays with a conventional spaced stereo microphone technique as illustrated in Figure 1.

In W-Ambisonics, two FOA microphones are spaced 17 cm apart to capture and represent interaural cues precisely, with two cardioid microphones spaced 50 cm apart, 50 cm in front, which improves frontal directionality. Varying the elevation of the cardioid pair enhances spatialization techniques such as reproduction with height channels. The combination of these two microphone pairs enables the translation of recorded audio into various reproduction formats according to

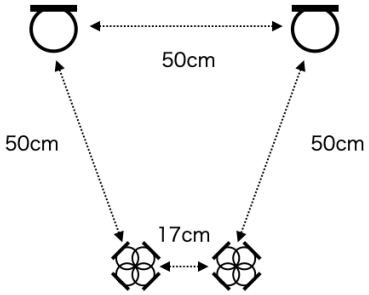

Fig. 1: “W-Ambisonics” microphone configuration. The cardioid pair is separated by 50 cm, located 50 cm behind two first-order Ambisonics microphones spaced 17 cm apart according to a standard interaural distance. Varying the height between pairs enables height spatialization.

practical limitations in reproduction peripherals. For example, the proposed method is scalable from 2-channel conventional stereophony (particularly spatially accurate binaural rendering) to 9-channel immersive audio. In this study, we tested subjective implications of the proposed method, to address the following research questions:

1. Can the proposed microphone technique provide better lateral localization than a conventional surround recording technique?
2. Can the proposed microphone technique be an efficient substitute for a conventional spaced array technique?
3. Is the reproduction of the proposed microphone technique “scalable”: can its signal be converted effectively across different playback systems, from binaural to 3D sound reproduction in multichannel loudspeaker arrays?

We designed and conducted three perceptual experiments to address these questions. First, we compared lateral auditory localization performance using audio recorded via the proposed method as compared to a conventional 5-channel surround microphone technique. Second, we devised a new binauralization method using the rear FOA microphone pair for headphone-based

reproduction, as each of two FOA microphones renders precise spherical harmonics analogous to the two ear positions of a human listener. Lastly, we made a solo piano recording using 5-channel surround technique, 7-channel immersive technique (5-channel plus two height channels), and the W-Ambisonic array, and subsequently evaluated the associated perceptual qualities of these three techniques.

2 Experiment 1: Lateral Localization Evaluation

As our initial experiment, we conducted a pilot study that compared human auditory localization performance using audio recorded via the proposed W-Ambisonics technique as compared to a conventional 5-channel recording technique. Our hypothesis was that W-Ambisonics enables improved performance in the lateral area. Instability in lateral localization is one of the most deficient issues for conventional 5-channel recordings [10, 11]. Only left-side localization was tested to simplify the experiment design.

2.1 Stimuli preparation

The experiment was conducted in the ECET Immersive Audio Lab at Rochester Institute of Technology (RIT). Seven Genelec 8020 loudspeakers were positioned as shown in Figure 2 as lateral sound sources. The loudspeakers were positioned at the azimuth angles from front left (45°) to rear left (135°) at 15-degree intervals. The distance between the center of the circle (listening position) and each loudspeaker was 320 cm.

Two sound sources (a pink noise pulse and a monophonic anechoic recording of xylophone) were reproduced via each of seven loudspeakers and captured using three microphone techniques as illustrated in Figure 2:

- One pair of cardioid microphones (DPA 4011), which represents the lateral segment of a surround recording microphone configuration (such as the Fukada Tree [12]).
- One cardioid microphone (DPA 4011) and one FOA microphone (Rode NT-SF1), together representing one of two lateral components of the W-Ambisonics technique. The FOA microphone was positioned at the center (listener ear) position, with the cardioid microphone 50 cm in front.

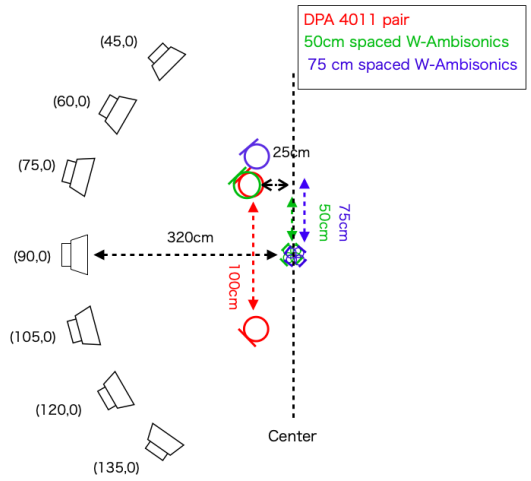

Fig. 2: Illustration of stimuli recording procedure. Seven loudspeakers were located in a semi-circular arc at 15° increments to create seven directional sound sources at 320 cm from the central recording position. These sources were recorded using three microphone techniques: the red color illustrates the pair of two cardioid microphones; green for the cardioid-FOA pair with a distance of 50 cm; blue for the cardioid-FOA pair with a distance of 75 cm.

- The same W-Ambisonics configuration as above, but with a 75 cm distance between microphones.

2.2 Listening test

For listening test playback, we used the same loudspeakers as used for stimuli recording, as shown in Figure 2. The seven stimuli recorded with the cardioid pair was reproduced through the front-left (45°) and rear-left loudspeaker (135°). The cardioid-FOA pair was reproduced through three loudspeakers: the cardioid microphone was reproduced through the same front-left loudspeaker, and the FOA microphone signals were decoded into the side-left (90°) and rear-left (135°) channel using the “Sampling or projection decoding (transpose)” method (*Higher-Order-Ambisonics-master* decoder [13]).

Participants listened to and reported the perceived image direction associated with the three recording techniques. We also included a non-processed, direct playback condition as the reference. It was assumed that

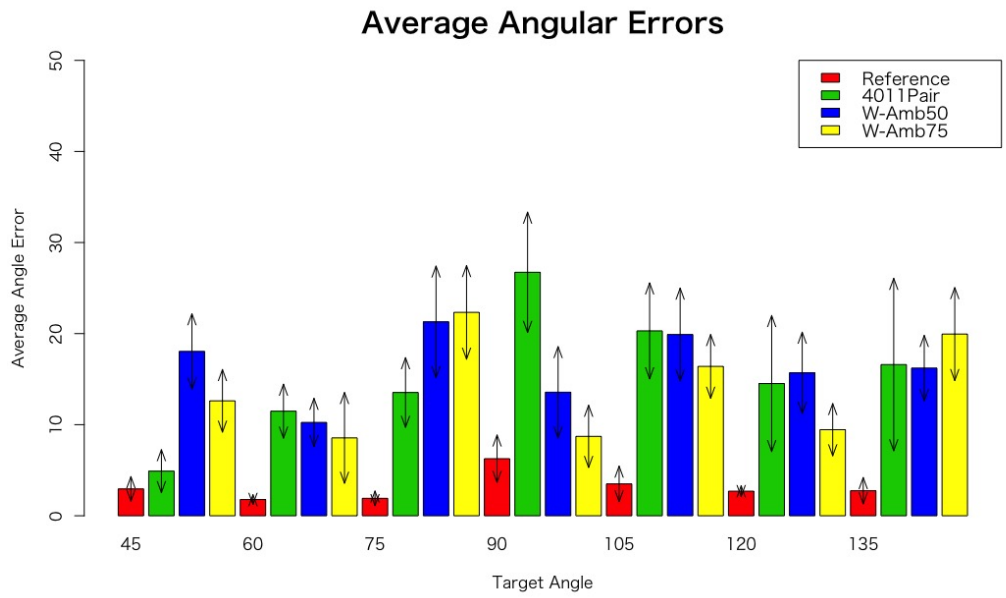

Fig. 3: Average angular errors in lateral auditory localization for seven source positions. Four colors represents different playback conditions (reference, cardioid pair, cardioid-FOA pair (50 cm), and cardioid-FOA pair (75 cm)) as indicated in the legend box, with double-ended arrows indicating the standard error (SE) of angular error for each position.

this reference would give the most accurate directional cue to the participants, so that we could compare localization performance across the three recording techniques. In total, each listener evaluated 56 stimuli (2 types of sound sources, 7 source positions, and 4 methods: cardioid pair, cardioid-FOA pair (50 cm), cardioid-FOA pair (75 cm), and reference (direct playback of source signals).

For reporting, participants indicated their perceived direction using a smart sensor, the Bosch-bno055 USB stick, which collected directional angles using the Bosch development desktop 2.0 software. Participants were seated in the center position, and the sensor was calibrated for each participant. The reproduction levels of all stimuli were matched at the listening position. The stimuli were presented in a randomized order, and the subjects could listen to a single stimulus repeatedly until they were able to report a perceived angle.

Six participants (RIT students and research staff) completed the listening test, who were were normal-hearing listeners and have experience with 3D sound, music production, and/or musical training.

2.3 Results

Figure 3 shows the average angular errors of reported localization angles. Compared to the reference stimuli (red bars), three methods produced about 18° error in average for lateral localization. This result is not surprising considering the limitation of a pairwise reproduction method, which would produce large angular errors and unstable images in the lateral area of a conventional surround reproduction system as previously reported in [10]. The interesting finding here is that the proposed W-Ambisonics method (in both 50 cm and 75 cm distances) afforded improved localization in the 90° direction. This indirectly supports that the proposed method is more effective in producing a reliable lateral image than a conventional 5-channel method. Readers should be advised that additional balancing (between the front, side, and rear channel) would be required to achieve the best reliable side image in a practical recording and mixing situation. Nevertheless, the proposed method produced enhanced localization of lateral sound images.

3 Experiment 2: Binaural rendering using two rear FOA microphones

As stated earlier, an FOA system can render sound sources’ directional information for all directions. Moreover, this characteristic is not deteriorated by a listener’s head movement, which is different from other microphone-based auralization methods. Therefore, it is possible to virtualize FOA loudspeakers for headphone reproduction with a head-movement-free sound field. Considering that the static binaural image has been a bottleneck for many headphone-based applications, this virtualization of an FOA sound field for headphone playback was a breakthrough for audio in virtual reality (VR) and augmented reality (AR) applications.

While FOA-based headphone sound rendering allows freedom of head movement, the perceived sound quality has not been optimal. Specifically, the spatial fidelity of a sound field from FOA is worse than that produced via conventional microphone techniques [14]. In addition, many researchers found that the FOA method can capture and render a small area of a sound field successfully, but not the entire sound field surrounding a listener’s head. This spatial inconsistency led us to hypothesize that two FOA microphones might sample and render the sound field correctly if located as analogous to the two ears of a human listener, towards ecological validity (realism). For this reason, our proposed method positions two FOA microphone arrays separated by the standard 17 cm interaural distance, as a human-centered generalization of sound field spatial sampling [15].

We conducted an informal listening test comparing a single FOA binauralized sound field with the proposed double-FOA method. The result shows that the double-FOA method provides listeners with a wide, realistic representation of musical instrument performance, which subsequently increased the overall sound quality. We are continuing our study and plan a future publication on findings.

4 Experiment 3: In-situ recording of a solo concert piano music

Based on our proposal and previous experiments, we designed and conducted a pilot experiment to test the W-Ambisonics technique for in-situ recording. For this experiment, we located both a conventional microphone array and the proposed W-Ambisonics array in McAfee concert hall of Belmont University in Nashville (TN, USA) and recorded a solo concert of piano music. The recordings were then mixed to (1) 5-channel surround (a conventional surround format), (2) 5-channel plus 2 height channels (a more immersive presentation through a vertical extension of the piano sound field), and (3) a 9-channel W-Ambisonics configuration (which can be considered equivalent to a standard 5-channel surround format, plus two additional horizontal side channels and two height channels). Additional side channels for (3) were added to support continuous lateral imaging based on results of Experiment 1. The three recording formats shared the same frontal (left, center and right) microphones and varied for rear and height channels. Figure 4 illustrates how these three recording methods were reproduced through loudspeakers. In this recording, we used the A-format signals of the FOA microphones as proposed in [9] to reduce any potential artifact from B-format conversion and maintain the best sound quality possible for high-fidelity music recording. We oriented each FOA microphone array so that three of its microphone

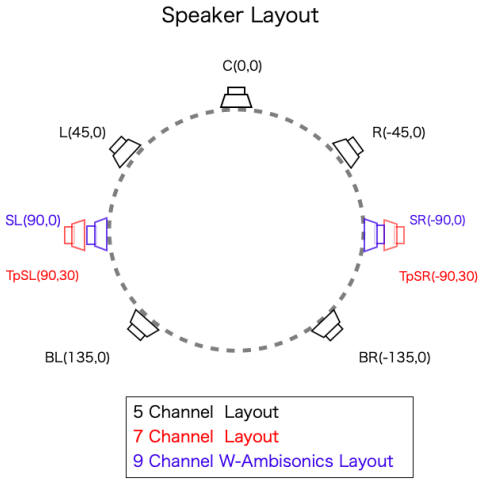

Fig. 4: Three playback configurations for immersive loudspeaker reproduction of the solo piano recording: (1) conventional 5-channel surround (black), (2) 5-channel surround plus two height channels (black + red), and (3) 5-channel surround plus two height and two additional side channels (black + red + blue).

capsules could face side, rear, and height directions to capture the musical performance stage. The capsules' signals were directly assigned to the corresponding loudspeakers for reproduction.

For our pilot listening study, we invited six audio professors and research assistants and evaluated their preferences regarding timbral fidelity and spatial fidelity. Participants were presented in random order three sound fields (5-channel surround, 7-channel surround+height, 9-channel front+FOA-decoded side, rear, and height) and they reported their perceived magnitudes using a 5-star Likert scale (worst, poor, fair, good, excellent). We asked participants to describe any idiosyncratic perceptual differences in comparing the three reproduced sound fields. We matched the loudness of the three sound fields using a NUENDO LUFS normalizer as well as by ear. The front channels of the 9-channel sound field sounded relatively quieter due to the presence of the many ambience components contributed by the FOA microphone arrays.

Analysis of the informal listening study showed a trend that the W-Ambisonics (9-channel surround) was evaluated either higher or similar to the 7-channel immersive representation of a solo concert piano. It is of interest (and of relevance to our ecological validity proposal) that overall preference was correlated with spatial fidelity. The participants appeared not to discriminate among the three reproduced sound fields based on spectral differences, but rather on spatial differences. One participant commented that the 9-channel sound field provided her with a more realistic concert hall experience yet she preferred the 7-channel version due to the clearer image of the piano. Her comments (along with similar opinions from other listeners) highlighted a plausibly relevant additional factor–the presence of visual cues. We hypothesized that overall preferences would be different if listeners could *see* the piano as well as the concert hall. To provide congruent visual immersion, we are preparing a new experiment that incorporates a 250° wraparound screen for immersive visuals with extended horizontal field-of-view (FOV) coverage (as illustrated in Figure 5). These planned experiments will explore the multimodal impact of visual congruency on perceived immersion and overall preference in virtual concert music experience.

Fig. 5: A three-dimensional (3D) visual experience of the grand piano in McAfee concert hall, simulated via a 250° wraparound screen, to be used in future congruent and holistic immersive listening tests.

5 Discussion

In some applications of sound field recording, it is necessary to achieve manageable portability of a microphone array and flexible scalability for multiple reproduction formats. We consider this portability and scalability as the underlying design criteria for the proposed W-Ambisonics technique. Whereas this technique may not produce the best sound quality for all applications, it can be used to record audio that translates across multiple reproduction formats ranging from binaural to a 9-channel format. In particular, the interaural spacing of the two FOA microphone arrays seems to render a convincing ambient sound field, which suggests a target application for using such a pair in field recording, and in contexts that prioritize accurate documentation of human listening perspectives, such as cultural heritage acoustics. Furthermore, with the recent growth in Virtual Reality technologies, manufacturers have begun producing affordable FOA audio recorders. Strategic use of these portable devices can be a powerful tool for the documentation and preservation of the aural heritage of cultural and natural sites, places of historical interest, and endangered locations around the world. Our study explores a novel configuration of extant tools in order to offer technical improvements to current practices and standards, enabling new applications in the arts and Humanities, social sciences, engineering and manufacturing sectors.

6 Summary

We propose and offer perceptual evaluations of a new portable 3D microphone technique, “W-Ambisonics,”

devised for human-centered virtual translation of acoustic events in real-world spaces. This technique combines a conventional stereo cardioid microphone array to capture the first wavefronts of sound sources with an interaurally spaced pair of FOA microphones to render the ambient sound field, thus centering these arrays in spatial analogy with human audition. Three listening experiments showed that the proposed microphone technique enhances lateral image precision, provides a wider binaural image than the single FOA method, and scales across multichannel reproduction formats. The scalable feature of the proposed method highlights its utility across fields, particularly for use in human-centered applications such as cultural heritage acoustics and soundscape research.

7 Acknowledgement

This study was supported by the National Endowment for the Humanities (NEH) project entitled Digital Preservation and Access to Aural Heritage Via a Scalable, Extensible Method (Award No. PR-263931-19).

References

[1] Gerzon, M. A., “Ambisonics in Multichannel Broadcasting and Video,” *Journal of Audio Engineering Society*, 33(11), pp. 859–871, 1985.

[2] Zotter, F. and Frank, M., *Ambisonics : A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality*, Springer Nature, 2019, doi:10.1007/978-3-030-17207-7.

[3] Cairns, P. and Moore, D., “Switched Spatial Impulse Response Convolution as an Ambisonic distance-panning function,” in S. Werner and S. Göring, editors, *ICSA 2019: Proceedings of the 5th International Conference on Spatial Audio*, pp. 99–106, Ilmenau Media Services, 2019, doi:10.22032/dbt.39961.

[4] Arteaga, D., “An Ambisonics Decoder for Irregular 3-D Loudspeaker Arrays,” in *Audio Engineering Society Convention 134*, 2013.

[5] Howie, W., King, R. L., Martin, D., and Grond, F., “Subjective Evaluation of Orchestral Music Recording Techniques for Three-Dimensional Audio,” in *Proc. Audio Engineering Society 142th Int. Conv.*, AES, Berlin,Germany, 2017.

[6] Wittek, H. and Theile, G., *Comparison of 7 surround main microphones*, http://www.hauptmikrofon.de/orf.htm, 2001.

[7] Kim, S., *Recording Piano in Surround: Discovering Preferences, Investigating Auditory Imagery, and Establishing Physical Predictors.*, Ph.D. thesis, McGill University, Montreal, Canada, 2009.

[8] Kamekawa, T. and Marui, A., “Evaluation of recording techniques for three-dimensional audio recordings: Comparison of listening impressions based on difference between listening positions and three recording techniques,” *Acoustical Science and Technology*, 41(1), pp. 260–268, 2020, doi:10.1250/ast.41.260.

[9] Dobson, A. and Woszczyk, W., “Tetrahedral Microphones: An Effective A/B Main System,” in *Audio Engineering Society Convention 147*, 2019.

[10] Martin, G., Woszczyk, W., Corey, J., and Quesnel, R., “Sound Source Localization in a Five-Channel Surround Sound Reproduction System,” in *Proc. Audio Engineering Society 107th Int. Conv.*, AES, New York, USA, 1999, preprint 4494.

[11] Kim, S., Ikeda, M., and Takahashi, A., “An optimized pair-wise constant power panning algorithm for stable lateral sound imagery in the 5.1 reproduction system,” in *Proc. Audio Engineering Society 125th Int. Conv.*, AES, San Francisco, USA, 2008.

[12] Fukada, A., “A challenge in multichannel music recording,” in *Proc. Audio Engineering Society 19th Int. Conf. on Surround Sound*, AES, Schloss Elmau, Germany, 2001.

[13] Politis, A., *Microphone array processing for parametric spatial audio techniques*, Doctoral thesis, School of Electrical Engineering, 2016.

[14] Martens, W. L. and Kim, S., “Relating Listener Preferences for Multichannel Sound Programs to Salient Auditory Attributes and Binaural Stimulus Measurements,” in *Proc. of Inter-Noise 2006*, INTER-NOISE, Honolulu, USA, 2006.

[15] Kolar, M. A., Ko, D., and Kim, S., “Preserving Human Perspectives in Cultural Heritage Acoustics: Distance Cues and Proxemics in Aural Heritage Fieldwork,” *Acoustics*, 3(1), pp. 156–176, 2021.