Audio Engineering Society

Convention Paper

Presented at the 151st Convention
2021 October, Las Vegas, NV, and Online

Implementing and Evaluating a Higher-order Ambisonic Sound System in a Multi-purpose Facility: A Lab Report

Sam Smith[1], W. Seth Helman[1], and Anıl Çamcı[1]

[1]*University of Michigan, School of Music, Theatre & Dance*

Correspondence should be addressed to Sam Smith (sdsmit@umich.edu)

ABSTRACT

Although Ambisonic sound reproduction has an extensive history, it started finding more widespread use in the past decade due to the advances in computer hardware that enable real-time encoding and decoding of Ambisonic sound fields, availability of user-friendly software that facilitate the rendering of such sound fields, and recent developments in immersive media technologies, such as AR and VR systems, that prompt new research into spatial audio. In this paper, we discuss the design, implementation, and evaluation of a third-order Ambisonic system in an academic facility that is built to serve a range of functions including instruction, research, and artistic performances. Due to the multi-purpose nature of this space, there are numerous limitations to consider when designing an Ambisonic sound system that can operate efficiently without interfering with the variety of activities regularly carried out in it. We discuss our approach to working around such limitations and evaluating the resulting system. To that end, we present a user study conducted to assess the performance of this system in terms of perceived spatial accuracy. Based on the growing number of such facilities around the world, we believe that the design and evaluation methods presented here can be of use in the implementation of spatial audio systems in similar multi-purpose environments.

1 Introduction

Ambisonics is a spatial audio technique that is rooted in research carried out in the 1970s [1]. It adopts a sound field approach to the capturing and reproduction of spatial audio scenes; this approach abstracts the relationship between the encoding of sound sources in a spatial context and their decoding to an output system, therefore making it possible to design or capture a scene once and diffuse it through different loudspeaker configurations. The scalability of this technique [2], paired with its ability to adapt to non-uniform loudspeaker layouts [3], has made it preferable among spatial audio researchers. Today, Ambisonics is finding wider-spread use as an intermediary format in modern immersive media applications such as VR and AR experiences.

In this paper, we evaluate a third-order Ambisonic system in the Chip Davis Technology Studio at the University of Michigan's School of Music, Theatre & Dance. The studio serves as a classroom, research laboratory and performance space. The acoustics and layout of

the room are designed to strike a balance between these functions, posing limitations to the implementation of an optimal sound system. We first provide details of the Chip Davis Technology Studio. We discuss our approach to design of the Ambisonic system and offer details of its implementation. We then describe a user study carried out to evaluate this system and report its results.

2 Related Work

One of the primary applications of the Ambisonic system discussed in this paper is artistic performance. The composer Ludger Brümmer argues that pieces that are composed and performed with high-density speaker arrays can offer more interesting, transparent, and multilayered sound settings [4]. Indeed, a 2011 survey by Peters et al. has demonstrated a rapidly growing interest among composers in the use of sound field techniques such as wave field synthesis and Ambisonics [5].

Numerous studies have investigated the design and implementation of Ambisonic loudspeaker systems (e.g., [6, 7, 8]). Furthermore, previous research has focused on the evaluation of such systems in controlled environments (e.g., [9, 10]). Power et al. carried out localization tests with a 16-speaker system in a semi-anechoic chamber, with 8 speakers positioned in a circle at ear level and 8 speakers positioned in a cube formation with speakers above and below the listener [11]. They compared first-, second-, and third-order Ambisonic reproduction using pink noise and speech signals. They found that while elevated sound sources with limited bandwidth could benefit from the increased precision of a higher-order system, lower-order reproduction can be sufficient for more diffuse sounds.

Kearney et al. evaluated listeners' perception of depth and distance in first- and higher-order Ambisonic sound fields reproduced over virtual loudspeakers (i.e., using Ambisonic-to-binaural decoding via headphones) [12, 13]. Their studies showed that the perception of depth and distance in a virtual Ambisonic space is not statistically different from the perception of depth and distance in a real acoustic environment regardless of the Ambisonic order.

Other aspects of Ambisonic sound reproduction have also been evaluated through user studies. Focusing on streamed spatial audio signals, Narbutt et al. proposed AMBIQUAL as a metric for the evaluation of localization accuracy and overall listening quality in first- and third-order Ambisonic sound fields rendered to binaural output [14]. The validation experiments that they carried out showed a strong correlation between listeners' subjective assessment of Ambisonic audio signals compressed for streaming and the AMBIQUAL analysis of the same signal. Their results also showed that lower bitrates had an adverse effect on the quality of the spatial audio experience.

Bertet et al. have evaluated the localization accuracy of sound fields encoded with Ambisonic microphones of various orders [15]. The authors used a ring of 48 speakers in an acoustically treated room, where 12 of the speakers were used for higher-order Ambisonic reproduction (i.e., virtual sources), whereas other speakers were used to play back the target sound, which was a 206-millisecond train of nine 22-millisecond white noise bursts. The participants were asked to use an acoustic pointer controlled with a dial to match the position of the targets. The pointer was a masking noise signal repeated in bursts of 150 milliseconds convolved with the encoded Ambisonic impulse responses for each control condition. The results revealed that localization accuracy was proportional to the order of the microphone and was affected by source incidence with lateral target directions at 75° and 120° yielding the highest confusion. Another study by Braun and Frank corroborated the correlation between Ambisonic order and localization [16].

In a qualitative analysis of spatial audio reproduction, Guastavino and Katz evaluated the performance of 1D (stereo), 2D (6-channel circular), and 3D (12-channel spherical) loudspeaker arrays in reproducing Ambisonic soundscape recordings [17]. Based on analyses of verbal responses and scale judgements gathered from two studies, the authors found a strong correlation between the source material and the optimal reproduction method. For instance, while the 3D configuration was found favorable for rendering indoor scenes, the 2D configuration was preferred for outdoor scenes. Although the listeners characterized the 2D setup as being more realistic than the 3D setup, authors suggest that a lack of familiarity with 3D sound reproduction systems might have prompted this result.

3 Chip Davis Technology Studio

The Ambisonic system at the core of this study is housed in the Chip Davis Technology Studio (referred

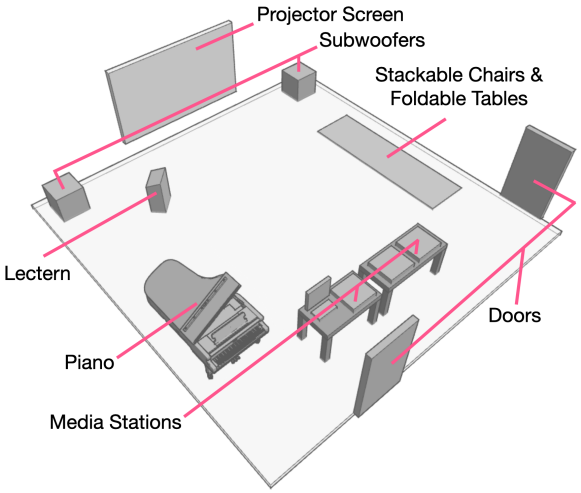

Fig. 1: Chip Davis Technology Studio floor plan.

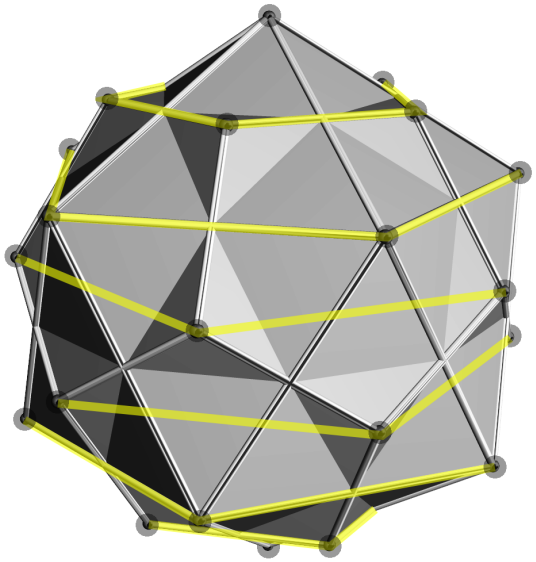

Fig. 2: A compound of two platonic solids, an icosahedron and a dodecahedron. The geometry has 32 vertices, indicated here with black dots. The yellow lines show the 6 pentagons, each of which spans 5 vertices with equal elevation.

to as Davis Studio hereafter), which is part of the Brehm Technology Suite at the University of Michigan's School of Music, Theatre & Dance. The Davis Studio is primarily used for the academic and artistic activities of the Department of Performing Arts Technology. Among these activities are course instruction, academic research, artistic performances and installations, seminars, and studio recording. The acoustics of the space are treated with various types of diffuser panels on the walls and ceiling, and bass traps in the corners; these treatments are designed to achieve a balance between the wet sound desired for concert purposes and the dry sound necessary for lectures. Due to the multi-purpose nature of this space, it consists of a host of music and media technologies. Besides the Ambisonic system, which we will detail shortly, the space is equipped with a 16-camera Qualysis motion capture system, an EON theatrical lighting grid, an HD projector and projection screen, a Roland M5000 mixer, a Bösendorfer CEUS grand piano, 10 folding tables, 40 stackable chairs, various computer hardware, a rolling white board and a lectern. While some of the equipment and furniture can be moved out of the Davis Studio as needed, there is often a compromise between the optimal states for different use cases at any given time. The studio is 30' long, 33'5" wide, and 12'5" high. Fig. 1, shows an overall floor plan of the studio.

4 Ambisonic System

The Ambisonic sound system in the Davis Studio is made up of 22 Genelec 8040B active loudspeakers. The speakers are mounted on the walls and ceiling using a custom scaffolding. The loudspeaker layout is modeled after another Ambisonic system we developed at the University of Illinois at Chicago's Electronic Visualization Lab; this system was based on a Platonic solid (i.e., convex regular polyhedron) geometry to ensure a regular speaker distribution with each speaker placed at a uniform distance from the origin (i.e., the sweet spot). Specifically, this geometry leveraged a combination of an icosahedron and a dodecahedron, as seen in Fig. 2, to achieve a 32-speaker configuration for fourth-order Ambisonic output. This geometry can be considered a midpoint triangulation of the icosahedron with a tessellation frequency of 2 [18], which allows the approximation of the icosahedron to a geodesic sphere with additional vertices. To maintain the optimal position of each speaker, a custom 16-feet scaffolding with an elevated floor was constructed. In the orientation shown in Fig. 2, the vertices of the compound geometry form 6 parallel pentagons; this way, the speakers could be mounted in rings of 5 speakers that have the same elevation with an additional speaker at the top of the geometry and another one at the bottom.

In the adaptation of this geometry to the Davis Studio,

one of the primary limitations was the need for the floor space of the studio to be kept accessible, meaning that the speakers could not be mounted in a spherical distribution around the sweet spot. Instead, they were projected from their optimal positions onto the walls and ceiling of the studio with the sweet spot taken as the origin of projection. Furthermore, mounting speakers below a certain height was avoided so as not to inhibit the movement of people and equipment in the studio. Finally, since the floor could not be elevated, it was not possible to mount a speaker under the sweet spot. As a result of these limitations, the fifth and sixth pentagons (i.e., the lowest two rings of 5 speakers) and the speaker at the bottom of the geometry had to be eliminated. Since one of the speakers in the fifth pentagon coincided with the area below the projection screen, it did not interfere with any foot traffic in the studio and could therefore be mounted. This resulted in a total of 22 speakers that are part of the eventual system.

A custom scaffolding was used on the walls and ceiling to mount the speakers in their calculated positions. The geometry shown in Fig. 2 simplified the scaffolding structure due to the matching elevation of the speakers on the top four pentagons. After the speakers were installed, a series of time-of-arrival and sound level measurements was carried out with impulse and continuous noise signals. These were used to apply delay and loudness compensations after the decoding stage to mitigate the distortions in the optimal geometry. The software for Ambisonic encoding and decoding is based on the ICST Ambisonics Externals for Max [19].

5 User Study

To evaluate the resulting Ambisonic system, we carried out a user study, where we assessed the accuracy of our system in rendering stationary and moving sound sources of different kinds at a constant distance with varying azimuth and elevation degrees. More specifically, we aimed to understand (1) how accurately the users can localize point sources rendered with our system, (2) whether the type of a sound source impacts this accuracy, and (3) whether moving sound sources are more easily localized than stationary sound sources.

5.1 Participants

The study was carried out with 20 participants, whose ages ranged between 18 and 24. None of the participants reported having hearing loss. The participants were recruited via email. 5 participants were Performing Arts Technology students, whereas the rest were University of Michigan students from various other majors.

5.2 Setup

We used the Qualysis motion capture system in the Davis Studio to track where the participants pointed at to indicate the perceived location of a sound source played back to them through the Ambisonic system. To achieve this, we equipped a 20-inch wooden dowel with 4 infrared markers, which were tracked by the motion capture system. Another set of markers were placed on a pair of lenseless goggles that the participants wore so that their head position could be tracked as a reference point. The tracking information was sent from Qualysis Track Manager (QTM) to Max via OpenSoundControl in real time. Here, the computed locations of the sound sources were stored with the motion tracking data pertaining to the participants' estimate of these locations.

5.3 Design and Stimuli

We used a between-subject design involving two trials. In the first trial, three types of stationary sound sources were used; these were music, speech, and impulse (i.e., brief noise burst repeated every 430 milliseconds). These sound types were chosen to cover the various use cases of the studio, wherein the audience is presented with audio material that displays diverse spectral, dynamic, and temporal qualities. The musical sound source, in particular, was a jazz track featuring guitar, drums, piano, vibraphone and double bass. Each sound type was presented to users in 24 different equidistant locations covering all areas of the Davis Studio, including locations above and below the ear level, and behind and in front of the participant. The distribution of these locations across the surface of a sphere can be seen in Fig. 3(a) in rectangular-projection view with 0° elevation corresponding to ear level and in Fig. 3(a) in overhead view with 0° corresponding to the front of the participant. Each sound type was presented in dedicated groups, but the ordering of the locations between groups were randomized within participant, and the ordering of the groups were randomized between participants to mitigate exposure bias. In the second trial, the music source from the first trial was presented on a trajectory moving around the participant in 3D at a constant distance. The participant was asked to follow the source by pointing at it with the dowel.

5.4 Procedure

After signing a consent form, the participant was given a description of the study. This was followed by a demonstration of the study apparatus, where one of the facilitators stood on the marked sweet spot and pointed at stationary and moving sound sources using the dowel with markers. Although egocentric pointing methods such as this one can impose a bias on the localization performance [20], we attempted to normalize this bias by emphasizing the optimal pointing technique in this demonstration. The participant was then asked to pick up the dowel and stand on the marked spot. They were instructed not to move their feet, legs, or torso substantially and to remain facing forwards during the study. In the first trial, once the facilitator presented a stationary sound source, the participant was asked to point at it to the best of their ability. When satisfied with their choice, they indicated this by saying the word, "good," at which point the facilitator stored the tracking information. This process was repeated for each of the 24 source locations for each sound type. In the second trial, the participant was presented with a moving musical sound source and was asked to track the source continuously with the dowel while their movement was sampled at 10 Hz. In a verbal survey following the trials, the participant was asked to comment on sound types and locations in terms of how easy or hard they were to localize during the study. They were also asked to compare the stationary and moving sound sources along the same criteria.

5.5 Results and Discussion

Using the Cartesian coordinates of the infrared markers sent from QTM to Max, a 3D vector that corresponds to the participant's pointing was calculated. The vector's intersection with the theoretical spherical surface where the sound sources were placed was derived in polar coordinates. Then, the great-circle (i.e., spherical) distance between the computational placement of a sound source and where a participant perceived it to be was calculated using the Haversine formula [21] seen below:

$$2r \cdot \arcsin\left(\sqrt{\sin^2\left(\frac{\varphi_2 - \varphi_1}{2}\right) + \cos\varphi_1 \cos\varphi_2 \sin^2\left(\frac{\lambda_2 - \lambda_1}{2}\right)}\right)$$

Box plots of these distances across all 20 participants for each of the 24 source locations and 3 sound types can be seen in Fig. 4(a, b, and c). Fig. 4(d) compares the medians from each of these box plots. Here, we can see that the sound type did not have a significant impact on localization accuracy, although the speech source seemed to yield a higher error in some cases. This might be due to the relative prominence of transient qualities in the music and impulse sources when compared to the speech source. This was also reflected in the survey responses, where 45% of the participant identified the speech source as the hardest to localize, whereas 55% of the participants considered the impulse to be the easiest to localize.

The median spherical distance between the computed and estimated stationary source locations averaged at 1.0157 radians on a scale between $0-\pi$ (SD: 0.4337, min: 0.3, max: 2.0). Sources below the ear level (#2, #8, #14, and #22) seem to have yielded a higher error. This can be explained by the performance of the Ambisonic system degrading due to the elimination of the bottom rings of speakers. The most accurately localized stationary sources were source #16, #3, and #4, which were all within an elevation of 15° and 50° above the ear level. Sources closer to the top (higher than 60° elevation) have also yielded lower than average median errors. The participants tended to perform better with sound sources towards the sides. The results show instances of front-back confusion, where a listener can perceive a source behind them as being in front of them and vice versa [22]. These confusions might be compounded by the floor and surface reflections in the studio. In the after-trial survey, the participants identified sources placed on the left and right sides of the room as the easiest to localize, while sounds directly behind and in front of them were found to be harder to localize.

In the moving source trial carried out with the music source, the median distance between the computed and estimated sources across all participants averaged at 0.4995 radians (SD: 0.136, min: 0.3, max: 0.9) on a scale between $0-\pi$. To compare this with the stationary source trial results, we focused on the music stationary sources that fall within the elevation range of the moving source (i.e., 26° – 54°). The average of the medians for qualifying stationary sources was 1.0157 radians. The improved performance in the moving source trial could be attributed to the role of a moving sound source in helping the listener resolve localization ambiguities [22], which primarily manifested in the stationary source trial in the form of front-back confusions.

Interestingly, 55% of the participants characterized the stationary sources as being easier to localize than the moving sound source.

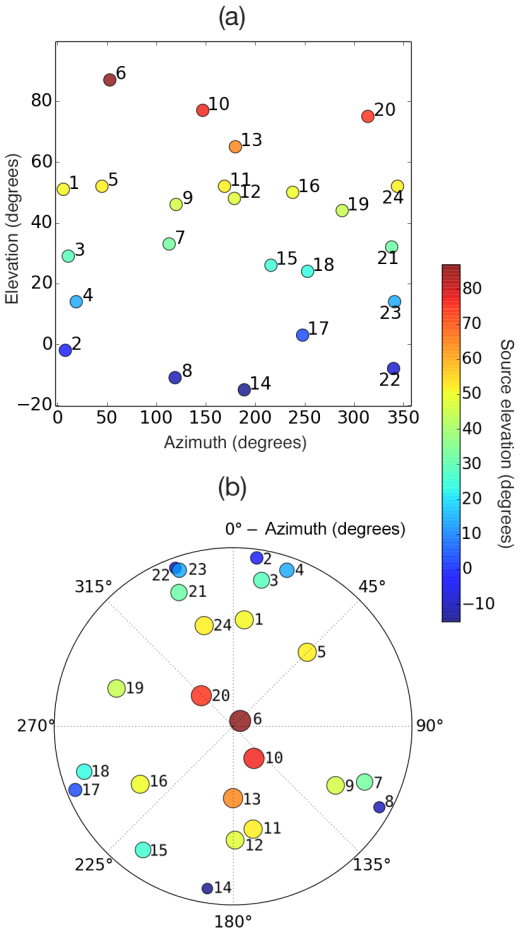

Fig. 3: Stationary source locations in the first trial. (a) Rectangular projection of the spherical surface where the sound sources were placed. (b) Overhead view of the sphere with elevation degrees encoded in color gradient and circle radius.

Fig. 5(a) shows a 2D projection of the 3D trajectory of the moving sound source, whereas Fig. 5(b) shows the corresponding plot of the median distance between computed and estimated sources over time. Although, the frontal region around ear level seems to have yielded relatively smaller median distances, the deviation in this region points to potential front-back confusions. At the extremes of the elevation range, the

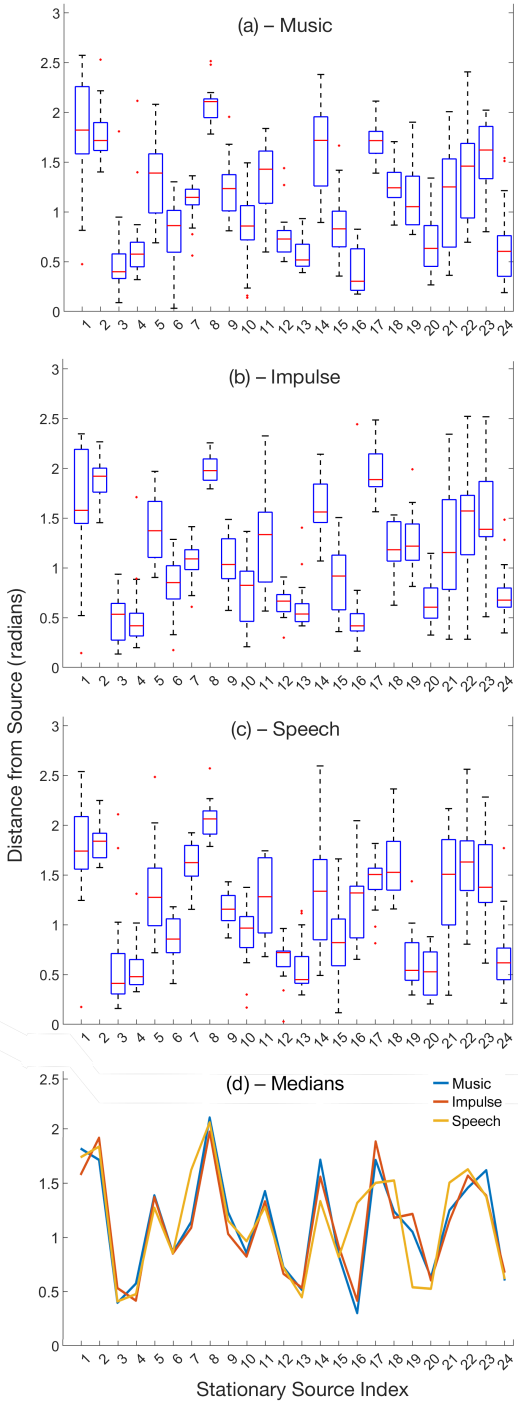

Fig. 4: Results of the trial with stationary sound sources. (a, b, c) Median distances between the 24 computed source locations and the participants' estimates for each sound type. (d) Median plots for each sound type.

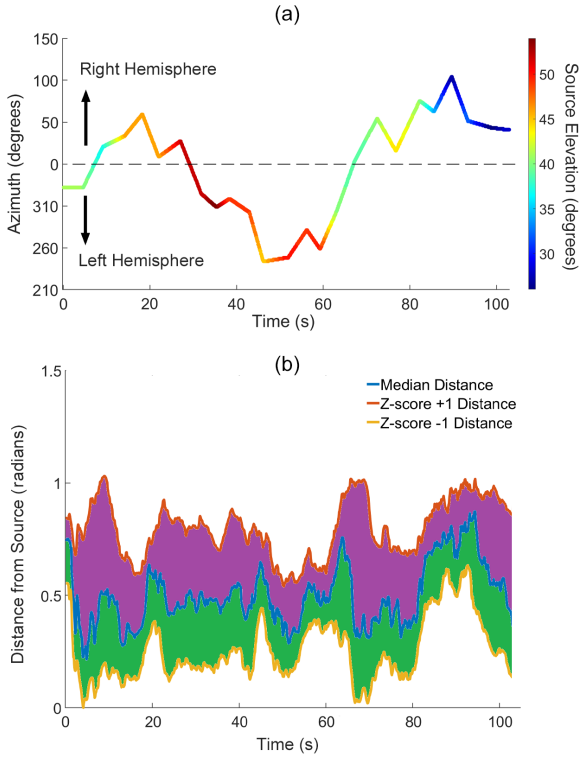

Fig. 5: Results of the trial with the moving sound source. (a) Source azimuth over time with source elevation indicated with a color gradient. (b) Median spherical distance between the computed source location and participants' estimation of this location.

median distances increase while the deviation seems to decrease, indicating that the error in these regions were consistent across participants. The greatest distance is observed at the lowest elevation value, which could be a byproduct of the lack of lower speaker rings and floor reflections in the studio. Although performance in the frontal and rear regions did not show significant differences, higher elevations in the rear region seem to have caused greater median distances than those in the frontal region.

6 Conclusion

The results of the study conform to our expectations for the most part. The Ambisonic system performed better where the speaker distribution was not affected by the placement limitations in the studio. The sound source type did not have a significant impact on localization although transient sounds may have helped. Outside of front-back confusions, that might be exacerbated by the reflections in the room, the participants performed well in pinpointing virtual sound sources, especially those that were positioned towards the left and right sides of the sound field. Furthermore, the participants performed better with the moving sound source, likely due to the constant change in the source position that may have helped with resolving localization ambiguities. These findings suggest that the system can adequately support artistic and academic applications in the studio although it may not be suitable for critical localization tasks that require high precision. In a follow-up study, we hope to investigate the accuracy of our system in rendering sources at varying distances. We also plan to evaluate how moving away from the sweet spot affects the localization accuracy. With recent developments in spatial audio and immersive media research, we expect that Ambisonic sound systems will continue to find increasing use in multi-purpose academic facilities such as ours. We hope that the implementation details of our Ambisonic system, the proposed method for its evaluation, and the results reported here will be useful to researchers who wish to implement similar systems.

References

[1] Gerzon, M. A., "Periphony: With-height sound reproduction," *Journal of the Audio Engineering Society*, 21(1), pp. 2–10, 1973.

[2] Frank, M. and Sontacchi, A., "Case study on ambisonics for multi-venue and multi-target concerts and broadcasts," *Journal of the Audio Engineering Society*, 65(9), pp. 749–756, 2017.

[3] Trevino, J., Okamoto, T., Iwaya, Y., and Suzuki, Y., "High order Ambisonic decoding method for irregular loudspeaker arrays," in *Proceedings of 20th International Congress on Acoustics*, pp. 23–27, 2010.

[4] Brümmer, L., "Composition and Perception in Spatial Audio," *Computer Music Journal*, 41(1), pp. 46–60, 2017.

[5] Peters, N., Marentakis, G., and McAdams, S., "Current technologies and compositional practices for spatialization: A qualitative and quantitative analysis," *Computer Music Journal*, 35(1), pp. 10–27, 2011.

[6] Färber, P. and Kocher, P., "The Mobile Ambisonics Equiptment Of The ICST." in *Proceedings of the International Computer Music Conference*, pp. 207–210, 2010.

[7] Ramakrishnan, C., Goßmann, J., and Brümmer, L., "The ZKM klangdom," in *Proceedings of the New Interfaces for Musical Expression Conference*, pp. 140–143, 2006.

[8] Lyon, E., Caulkins, T., Blount, D., Ico Bukvic, I., Nichols, C., Roan, M., and Upthegrove, T., "Genesis of the cube: The design and deployment of an hdla-based performance and research facility," *Computer Music Journal*, 40(4), pp. 62–78, 2016.

[9] Pulkki, V. and Hirvonen, T., "Localization of virtual sources in multichannel audio reproduction," *IEEE Transactions on Speech and Audio Processing*, 13(1), pp. 105–119, 2004.

[10] Lopez-Lezcano, F. and Jette, C., "Bringing the GRAIL to the CCRMA Stage," in *Proceedings of the Linux Audio Conference*, 2019.

[11] Power, P., Davies, W., Hirst, J., Dunn, C., et al., "Localisation of elevated virtual sources in higher order ambisonic sound fields," *Proceedings of the Institute of Acoustics*, 2012.

[12] Kearney, G., Gorzel, M., Boland, F., and Rice, H., "Depth perception in interactive virtual acoustic environments using higher order ambisonic soundfields," in *Proccedings of the 2nd International Symposium on Ambisonics and Spherical Acoustics*, 2010.

[13] Kearney, G., Gorzel, M., Rice, H., and Boland, F., "Distance perception in interactive virtual acoustic environments using first and higher order ambisonic sound fields," *Acta Acustica united with Acustica*, 98(1), pp. 61–71, 2012.

[14] Narbutt, M., Allen, A., Skoglund, J., Chinen, M., and Hines, A., "AMBIQUAL-a full reference objective quality metric for ambisonic spatial audio," in *Proceedings of the 10th International Conference on Quality of Multimedia Experience*, pp. 1–6, IEEE, 2018.

[15] Bertet, S., Daniel, J., Parizet, E., and Warusfel, O., "Investigation on Localisation Accuracy for First and Higher Order Ambisonics Reproduced Sound Sources," *Acta Acustica united with Acustica*, 99, p. 642 – 657, 2013.

[16] Braun, S. and Frank, M., "Localization of 3D ambisonic recordings and ambisonic virtual sources," in *Proceedings of the International Conference on Spatial Audio*, 2011.

[17] Guastavino, C. and Katz, B. F., "Perceptual evaluation of multi-dimensional spatial audio reproduction," *The Journal of the Acoustical Society of America*, 116(2), pp. 1105–1115, 2004.

[18] Hollerweger, F., *Periphonic sound spatialization in multi-user virtual environments*, Master's Thesis, Austrian Institute of Electronic Music and Acoustics (IEM), 2006.

[19] Schacher, J. C., "Seven years of ICST Ambisonics tools for maxmsp–a brief report," in *Proc. of the 2nd International Symposium on Ambisonics and Spherical Acoustics*, 2010.

[20] Bahu, H., Carpentier, T., Noisternig, M., and Warusfel, O., "Comparison of different egocentric pointing methods for 3D sound localization experiments," *Acta acustica united with Acustica*, 102(1), pp. 107–118, 2016.

[21] Chopde, N. R. and Nichat, M., "Landmark based shortest path detection by using A* and Haversine formula," *International Journal of Innovative Research in Computer and Communication Engineering*, 1(2), pp. 298–302, 2013.

[22] Wightman, F. L. and Kistler, D. J., "Resolution of front–back ambiguity in spatial hearing by listener and source movement," *The Journal of the Acoustical Society of America*, 105(5), pp. 2841–2853, 1999.