Audio Engineering Society

Convention Paper 10455

Presented at the 150th Convention
2021 May 25-28, Online

Comparing immersive sound capture techniques optimized for acoustic music recording through binaural reproduction

Will Howie[1], Denis Martin[2], Toru Kamekawa[3], Jack Kelly[2], and Richard King[2]

[1] *CBC/Radio–Canada, Vancouver, BC, Canada*

[2] *McGill University, Graduate Program in Sound Recording, Montréal, QC, Canada*

[3] *Tokyo University of the Arts, Department of Musical Creativity and the Environment, Tokyo, Japan*

Correspondence should be addressed to Will Howie (wghowie@gmail.com)

ABSTRACT

A study was undertaken to compare three immersive sound capture techniques optimized for acoustic music recording, within the context of binaural audio reproduction. 3D audio stimuli derived from 9-channel (4+5+0) recordings of a solo piano were binaurally rendered and presented to listeners over headphones. Subjects compared these stimuli in terms of several salient perceptual auditory attributes. Results of the double-blind listening test found no significant differences between two of the sound capture techniques, "spaced" and "near-coincident," for the perceptual auditory attributes "envelopment," "naturalness of sound scene," and "naturalness of timbre." The spaced technique, however, was shown to create a larger virtual image of the sound source than the near-coincident technique. The coincident technique was found to create an immersive sound scene that occupies a different perceptual space from the other two techniques, delivering less envelopment and naturalness.

1 Introduction

The previous decade has seen a proliferation of research, development, and commercialization of immersive media systems. These range from immersive broadcast formats such as Japan Broadcasting Corp.'s Super Hi-Vision [1], to 3D films paired with immersive audio formats such as Dolby Atmos or Auro 3D, to virtual and augmented reality systems using various types of head-mounted displays [2], to mixed immersive/interactive media installations in art galleries and industrial spaces[1]. To successfully bring the user into a new augmented/merged reality, not only does the visual aspect of the virtual scene need to be convincing, but also the sonic aspect. The COVID-19 pandemic has shown that there exists a large appetite for streaming audio/video immersive media content, particular presentations of live music.

A number of immersive audio formats have been introduced and standardized, [7] and numerous music recording techniques for said systems have also been discussed, particularly for acoustic music [3, 4]. Acoustic music (e.g., orchestral, chamber, choral, etc.) plays a large role in film and video game scores, live/streaming/broadcast concert performances, music production, and can be featured in any number of compelling virtual reality (VR) or augmented reality (AR) experiences. Immersive sound capture techniques optimized for acoustic music have been compared in previous studies (see: Section 1.2), but almost entirely through loudspeaker-based sound reproduction. Given the complexity and cost of installing even the simplest immersive audio system in the home environment,

[1] TeamLab's large-scale installation/museum in Odaiba, Tokyo, Japan (https://borderless.teamlab.art) exemplifies what is possible in terms of mixing visual art with interactive/immersive environments.

one can expect many virtual/immersive music experiences will be realized through binaural reproduction over headphones. This will almost certainly be the case for VR/AR experiences that rely on the use of head-mounted displays. As such, it would be valuable to understand how different acoustic music immersive sound capture techniques compare within the context of binaural sound reproduction.

1.1 3D sound capture techniques for acoustic music reproduction

In recent years, numerous sound capture techniques for three-dimensional (3D) reproduction of acoustic music have been developed and introduced [3–6, 31]. These sound capture techniques have typically been optimized for reproduction using standardized loudspeaker-based 3D audio formats (see: [7]), ranging in size and complexity from 4+5+0 (five loudspeakers at ear-level; four elevated loudspeakers) to 9+10+3 (ten loudspeakers at ear-level; nine elevated loudspeakers; three loudspeakers at floor-level). These techniques can generally be divided into three categories:

1.1.1 Spaced techniques

Spaced techniques capture and reproduce spatial sound information through time of arrival differences between microphone signals. A one-to-one microphone signal to loudspeaker relationship is typically maintained (e.g., 9 microphones for 9 loudspeakers). Many proposed spaced 3D sound capture techniques emphasize distant spacing between side, rear, and height microphones to ensure effective decorrelation between microphone signals [3–6], which is believed to be important for reproducing enveloping ambient sound fields [8–10].

1.1.2 Near-coincident techniques

Near-coincident techniques capture and reproduce spatial sound information through a combination of timing and level differences between microphone signals. Smaller spacing between microphone capsules are used as compared with spaced techniques: typically less than 1m. Lee [11] and Wallis and Lee [12] have provided suggestions for microphone polar patterns and angles to avoid vertical inter-channel crosstalk when positioning near-coincident height layer microphones. Again, a one-to-one microphone signal to loudspeaker relationship is typically maintained.

1.1.3 Coincident techniques

Coincident techniques use only intensity differences between microphone signals to capture and reproduce spatial sound information. As implied by the name, microphone capsules should be positioned as close to physically coincident as possible. Coincident immersive recording techniques are generally considered to be channel-number or format agnostic; microphone signals therefore require matrixing or post-processing to achieve correct decoding for a given reproduction system. These techniques are often discussed within the context of ambisonics [13–15], a specific approach to sound field capture and reproduction introduced by Gerzon [16].

1.2 Comparing 3D sound capture techniques for acoustic music

Several previous studies have compared 3D sound capture techniques optimized for acoustic music reproduction. Ryaboy [15] investigated perceptual differences between a coincident technique, “MS+Z”, and a mixed spaced/coincident technique, “Twins Square,” and found significant differences between the two in terms of horizontal and vertical sound source localization, and perceived room size. In two separate studies comparing 3D sound capture techniques for acoustic music reproduction, Howie and his co-authors [10, 17] found clear perceptual differences between the various techniques under investigation. In both studies, techniques within the same design family (e.g., “spaced techniques”) tended to be rated similarly, across a wide range of attributes. Also, spaced and near-coincident techniques appeared to occupy a more common perceptual space as compared with coincident techniques. Kamekawa and Marui [18] also compared spaced, near-coincident, and coincident sound capture techniques within the context of acoustic music recording. They found that the spaced technique was the most robust in terms of retaining sound scene impression across different listening positions. In that study, listeners tended to perceive the three techniques thusly: "It is also estimated that Ambisonics gives the impression of 'hard,' the near-coincident array gives 'rich' and 'wide,' and Spaced Array gives 'clear' and 'presence.' [18, p.268]"

1.3 Comparing 3D sound capture techniques through binaural reproduction

The studies discussed in section 1.2 all took place within the context of loudspeaker-based sound reproduction. As already discussed, it is likely that many consumers will experience 3D audio through headphone-based binaural sound reproduction. Surprisingly then, almost no formal studies comparing immersive sound capture techniques through binaural rendering have included music recordings as stimuli. Millns and Lee [19] created room impulse responses of three different 360° microphone arrays: Equal Segment Microphone Array (ESMA), First Order Ambisonics, and a Neumann KU-100 dummy head microphone. Music signals were convolved through each array. Listeners found the KU-100 and ESMA techniques tended to produce wider source images and wider environmental images than the ambisonics technique.

1.4 Goals of the current study

The studies discussed in section 1.2 all found that within the context of loudspeaker-based reproduction, clear perceptual differences can be observed between spaced, near-coincident, and coincident 3D sound capture systems optimized for acoustic music recording and reproduction. The goal of the current study is to compare a representative 3D microphone array from each of those three categories through binaural reproduction, and investigate what differences listeners may perceive between the three techniques. Although it may seem counterintuitive to compare channel-specific with channel-agnostic recording techniques, this is representative of not only many previous studies comparing multichannel sound capture systems [10, 15, 17–22], but also the kind of real-world comparisons regularly undertaken by professional recording engineers.

2 Method

2.1 Sound capture techniques under investigation

A previous study by Howie et al. [10] compared four different 3D microphone arrays optimized for music recording and reproduction for 4+5+0. For the current study, the stimuli from three of those techniques were binaurally rendered. The fourth technique was omitted from the current study as it was found in [10] to be perceptually very similar to the included "spaced" technique. A short summary of each technique used in the current study follows:

2.1.1 Spaced technique – Technique 1

This technique, designed by Howie and described in detail in [6] and [10], uses an array of three spaced omni-directional microphones to capture primarily direct sound, and an array of six widely spaced directional microphones (cardioid and wide-cardioid) to capture spatially diffuse ambience.

2.1.2 Near-coincident technique – Technique 2

This technique was designed by Theile and Wittek, and described in detail in [23]. The technique uses shorter spacings between its nine microphones than the spaced technique. Microphones with polar patterns of greater directivity (cardioid and super-cardioid) are prescribed to ensure adequate channel separation.

2.1.3 Coincident technique – Technique 3

Designed by Geluso, and described in detail in [14] and [15], this technique is comprised of a front-facing cardioid microphone capsule, a rear-facing cardioid capsule, a laterally oriented bi-directional capsule, and a vertically oriented bi-directional capsule, all placed as close to physically coincident as possible. Geluso provides a detailed scheme for signal matrixing to achieve a 4+5+0 compliant mix. This technique can also be considered a "native b-format" capture system: the necessary W, X, Y, and Z signals can be derived from combinations and subtractions of the various microphone signals.

2.2 Production of stimuli

2.2.1 Stimuli recording

The three sound capture techniques under investigation were setup for simultaneous recording of a solo piano (Figure 1). The choice of a piano as sound source was based on its complex radiation patterns, large timbral and harmonic range, ubiquity within many genres of music, and familiarity with subjects who have been trained in an academic music program. The recording venue was the Music Multimedia Room at McGill University (Figure 1). At the time of recording, this large (24.4 m x 18.3 m x 17 m) scoring stage contained no acoustical treatment: RT60 measured approximately 4.5 s. Microphone choice and placement for each technique was based on specifications and recommendations from their creators. Final placement of microphones was optimized by two professional recording engineers, both of whom had significant experience recording and mixing immersive audio. A detailed list of microphones used can be found in [10]. All microphones were routed to a Sony SIU-100 System Interface Unit, using the internal microphone preamps and A/D converters. Recordings were made to a Pyramix workstation at 96 kHz / 24 bit resolution. Monitoring and mixing took place at McGill University's Studio 22, which is equipped with 28 full-range, two-way loudspeakers (ME Geithain *M-25*) powered by Flying Mole class D amplifiers. The loudspeakers are arranged for reproduction of both 9+10+3 and 4+5+0, as per International Telecommunications Union guidelines [7].

2.2.2 Stimuli mixing

A 25 s excerpt of J. S. Bach's "Variation 13" from the *Goldberg Variations* was chosen as stimulus. Each technique was balanced by a team of two professional recording engineers, both having over ten years of experience recording and mixing multichannel audio, one of whom was also the principal recording engineer for the stimuli. Attention was given to maintaining a similar balance of direct to reverberant sound for each technique. Microphone signals were not filtered in any way. Matrixing of the coincident technique's signals to achieve a 4+5+0 rendering strictly followed creator guidelines [14]. This mixing process resulted in three stimuli, each corresponding to one of the techniques under investigation. More details on the recording, mixing, and level-matching of the original loudspeaker-based stimuli can be found in [10].

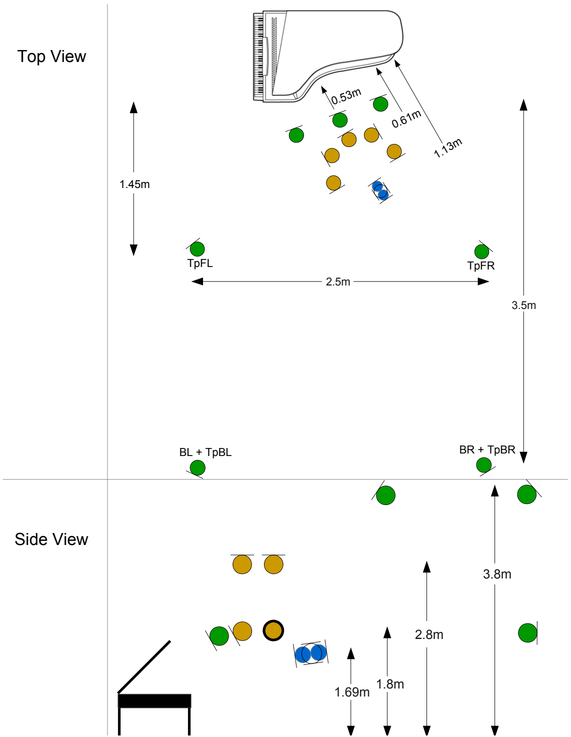

Figure 1. Top-down and side view of recording techniques under investigation. Technique 1 in Green, Technique 2 in Yellow, Technique 3 in Blue.

2.2.3 Stimuli binaural rendering and level matching

The three 9-channel stimuli were rendered for binaural reproduction within the spatial audio software "SPAT Revolution", using the built-in "KEMAR 1" head-related transfer function (HRTF) model. Several commercial, free-ware, and custom-built binaural rendering plugins and software were auditioned before deciding on this solution, which was found to perform best in terms of minimizing timbral changes within the stimuli, when compared with loudspeaker reproduction. The decision to use a generalized HRTF model was based on the impracticality of measuring individualized HRTFs and binaural room impulse responses for each subject, a method that is also not representative of current immersive virtual content creation or distribution.

The parameters within the binaural panning model were set to match loudspeaker azimuth, angle of elevation, and distance from the listener within Studio 22, where the stimuli were originally mixed. The rendering was then checked by the principal recording and mixing engineer of the stimuli. Based on their recommendations, small modifications were made to the binaural rendering model in terms of the distance of the virtual speakers from the listener. This was found to give an impression of the stimuli that better matched with the loudspeaker-based reproduction. The same rendering model was used for all three stimuli. Individual channel levels within the original 9-channel mixes were not altered. Audio resolution remained at 96 kHz / 24 bit. The binaural rendering process was monitored using Sennheiser HD650 headphones. Integrated loudness measurements were performed on the resultant stereo binaural stimuli using a professional software loudness meter calibrated to the EBU+9 scale [24]. Gain adjustments were made to the audio files until all three stimuli measured to within 0.1 LUFS of each other.

2.3 Subjective evaluation of stimuli

A double-blind listening test was designed to identify possible perceptual differences between the three 3D sound capture techniques under investigation.

2.3.1 Spatial audio perceptual attributes

Four perceptual attributes were chosen for investigation: “envelopment”, “naturalness of sound scene”, “naturalness of timbre”, and “sound source image size”; definitions can be found in [10]. These attributes were arrived at by a panel of four professional recording engineers/audio researchers who were asked to compare stimuli corresponding to the original four sound capture techniques from [10] in an informal listening session, and then determine the sound attributes that best quantified perceived differences between the techniques.

2.3.2 Testing venues

The listening test took place at two different locations: McGill University’s Schulich School of Music and Tokyo University of the Arts’ Department of Musical Creative and the Environment. The testing venue at each location was an acoustically isolated room. For both venues, Sennheiser HD650 headphones were used in conjunction with an RME Fireface UFX+ audio interface. This model of open headphones is the same that was used while monitoring the creation of the binaural stimuli. Head-tracking was not a component within the audio playback, as it was not possible to implement within the equipment and software available to the researchers.

2.3.3 Implementation of listening test

The listening test was implemented using Cycling 74’s Max/MSP software. Each subject was seated in the testing venue, explained the testing conditions, and given time to familiarize themselves with the interface and stimuli. Definitions of perceptual attributes were provided in English, both verbally and in written form. For each trial, subjects were asked to evaluate stimuli labelled “A”, “B”, and “C” for a given attribute, using a set of continuous sliders (0-100). Each letter corresponded to one of the three techniques under investigation. Anchor words were provided at the extremes of each slider. Since absolute anchors were not given at intervals along the scales, these measurements are relative and not absolute. To reduce scaling bias, subjects were instructed to rate the stimulus they felt was the “most” or “best” of a given attribute as 100%, then using that as a reference, rate the other two accordingly. More than one stimulus could be rated as 100%. Subjects were also instructed to treat each trial as a “new test”, and not attempt to base their ratings on their memory of responses from previous trials. Subjects completed four trials per attribute, for a total of 16 trials.

Subjects could switch between playback of A, B, and C stimuli or stop playback at any point. Playback of stimuli was continuously looped and time synced to ensure seamless switching. For each trial, stimulus assignments to A, B, and C, and the perceptual attribute to be rated were randomized within the testing program. Subjects were instructed to set a comfortable listening level before completing the first trial, then leave the level unchanged for the remainder of the test. At the test’s midway point was an enforced rest period of 1 minute. Subjects took an average of 25 minutes to complete the test, after which they were asked to complete a short demographic survey. Investigators were not present in the venue during testing.

2.3.4 Subjects

A total of 19 subjects across both venues participated in the listening test. Subjects were selected based on Howie et al.'s [25] findings regarding the most valuable types of previous experience for listener consistency within 3D audio evaluation. All subjects had at least four years music recording/production experience and musical training, with the averages across participants being 9.1 years and 14 years respectively. The average age of subjects was 29 years old. Previous experience listening to binaural renderings of 3D music recordings was almost universal.

3 Results

3.1 Z-scores

Attribute ratings collected from the participants were relative and not absolute. Therefore, the responses were converted to z-scores, also known as scaling for mean and standard deviation, to normalize for overall differences in the way participants used the rating scale (as seen in Figure 2). The data was normalized for mean and standard deviation for each participant within each attribute, as per International Telecommunications Union recommendations [26].

3.2 Attribute ratings

The average rating for each recording technique for each attribute is visualized in Figure 2.

A one-way ANOVA and subsequent pairwise t-tests with the Bonferroni correction were performed for each attribute. The results are displayed in Tables 1 and 2.

The one-way ANOVA shows a main effect of "array" for all attributes. Further analysis with the pairwise t-tests revealed that the coincident technique was rated significantly lower than the spaced and near-coincident techniques for all auditory attributes under investigation. The spaced and near-coincident techniques were not rated significantly differently for any attribute except for "sound source image size": the spaced technique was rated as creating a "larger" perceptual sound source between the two techniques.

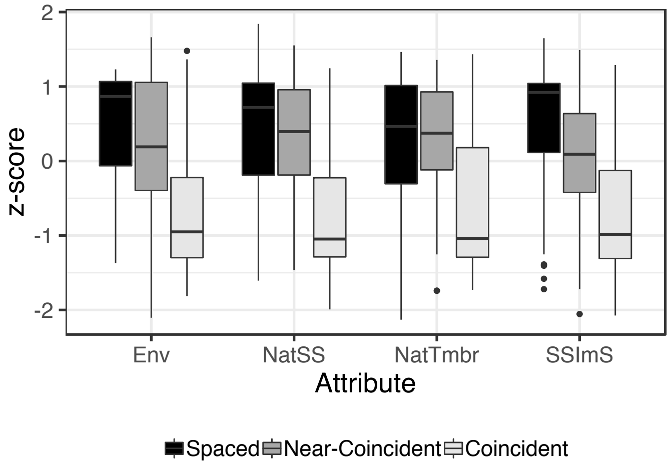

Figure 2. Attribute ratings averaged across all participants for each recording technique. Abbreviations: "Env" = envelopment, "NatSS" = naturalness of sound scene, "NatTmbr" = naturalness of timbre, "SSImS" = sound source image size.

Attribute	**Spaced Mean**	**Near-coincident Mean**	**Coincident Mean**	**F (df)**	***p***
Env	0.47	0.18	-0.65	38.58 (2,237)	<.001
NatSS	0.37	0.32	-0.68	40.79 (2,237)	<.001
NatTmbr	0.27	0.29	-0.56	24.97 (2, 237)	<.001
SSImS	0.54	0.06	-0.61	38.02 (2, 237)	<.001

Table 1. One-way ANOVA results for each attribute. Mean scores shown are from the transformed data (z-scores), as explained in Section 3.1.

	Spaced	Near-coincident
Envelopment		
Near-coincident	.03	---
Coincident	**<.001**	**<.001**
Naturalness of Sound Scene		
Near-coincident	.71	---
Coincident	**<.001**	**<.001**
Naturalness of Timbre		
Near-coincident	.87	---
Coincident	**<.001**	**<.001**
Sound Source Image Size		
Near-coincident	**<.001**	---
Coincident	**<.001**	**<.001**

Table 2. Pairwise t-test results (*p*). Statistically significant results displayed in **bold**.

4 Discussion

4.1 Perceptual differences between sound capture techniques

As can be seen in Figure 2, and Tables 1 and 2, listeners in the current study observed clear perceptual differences between the spaced and coincident techniques, and the near-coincident and coincident techniques. Or, put another way, the spaced and near-coincident techniques occupied a similar perceptual space as compared with the coincident technique. In two previous studies by Howie et al. [17] and Kim et al. [27], listeners found two spaced 3D microphone arrays to be perceptually similar as compared with a coincident array, specifically a spherical array designed for higher order ambisonics sound capture. Similarly, in [10], the loudspeaker-based study from which the current study's stimuli-set is derived, listeners rated the two spaced recording techniques nearly identically, and the near-coincident technique similarly to the spaced techniques. In contrast, the attribute ratings for the coincident technique were distinctly separate from the other techniques under investigation. In the current study, the coincident technique gave listeners an acoustic music sound scene that lacks in envelopment and naturalness as compared with the other two techniques, and also delivers a smaller image of the piano. Millns and Lee [19] found that the ESMA technique (a compact, near-coincident technique) tended to produce wider source images and wider environmental images than a coincident technique under investigation.

Techniques 1 and 2 are both designed with channel separation/decorrelation in mind: Technique 1 aims to achieve this through large physical spacing between microphones, and Technique 2 through microphone polar patterns of greater directivity. Decorrelation of reflected energy at the two ears is known to be an important factor in two different acoustical measures of spaciousness: listener envelopment (LEV) and apparent source width (ASW) [28], which are related to the perceptual attributes "envelopment" and "sound source image size" in the current study. By design, the sound fields reproduced by Techniques 1 and 2 should contain a higher degree of decorrelation between microphone signals than for Technique 3. This may explain the stark difference in ratings observed between the techniques for those two attributes. This is line with Griesinger's [8] assertion that decorrelation of the ambient component of recordings across the audible frequency spectrum is necessary for achieving optimal levels of "spaciousness" in reproduced sound.

In this study, there were no significant differences between the spaced and near-coincident techniques for the perceptual attributes "envelopment," "naturalness of sound scene," and "naturalness of timbre". For the attribute "sound source image size," the spaced technique was rated significantly higher than the near-coincident technique. This result is in contrast to Kamekawa and Marui's recent study [18], in which the near-coincident technique was observed as giving a "wide" image. This discrepancy is likely due to the differences in design philosophy between the near-coincident techniques used in [18] and the current study. For the near-coincident technique used in this study, the frontal sound image is captured by a combination of a front-facing centre cardioid microphone, and laterally-facing left and right super-cardioid microphones. With this arrangement, the left and right microphones will naturally capture less frontal direct sound than the mono centre microphone, which may result in a somewhat narrower sound source image as compared with the spaced technique, depending on how the techniques have been optimized or mixed. This was also observed in [10], where the two spaced techniques were rated as producing a larger sound source image size than the near-coincident technique. This is not to suggest that a "wider" or "larger" sound source image is preferred: optimal image size in acoustic music recordings is informed by myriad artistic and aesthetic considerations.

4.2 Practical considerations for 3D acoustic music recording techniques when creating content for virtual environments

For those wishing to capture and reproduce acoustic music within virtual or augmented spaces, there are several considerations that, in some ways, supersede the issues of perceptual space discussed in Section 4.1. Firstly, there is the question of perspective. Most spaced techniques, as well as a number of near-coincident techniques, are designed with a "concert" or "cinematic" perspective in mind: music in front, ambience from the sides, rear, and above. Such techniques can deliver high degrees of envelopment or presence, but may not be well-suited to musical sound scenes that incorporate a more 360° or "wrap-around" perspective; the techniques are not designed to prioritize accurate localization of direct sound sources at the sides or rear of the listener. For acoustic music sound scenes where the listener is to be surrounded by the performers, coincident techniques, or techniques based on equal spacing/segmentation of microphones would likely better recreate the sound scene in terms of true image localization. Lee's Equal Segment Microphone Array (ESMA) [29] and Wittek and Theile's ORTF-3D [30] are two examples of such an approach.

Another important consideration when capturing immersive audio of any kind is the complexity of setup for a given microphone array. Although it involved the largest spacing between microphones, and most microphone stands, in the current study the authors found Technique 1 to be both the fastest to setup, and optimize. This may be due to there being no specific distance prescribed between microphones – the recording engineer can therefore quickly position microphones based on what they are hearing within the recording venue. For techniques that make use of specific microphone angles and distances, such as Technique 2, or ESMA and ORTF-3D, the use of a purpose-built microphone array rig/stand would vastly reduce setup time. A self-contained, single-point, coincident recording system, such as a tetrahedral or spherical microphone array, would certainly be the fastest to setup, though optimizing their placement within the recording venue can be difficult if correct, real-time matrixing or rendering of their signals is not possible.

Finally, one must also consider whether or not video, especially 360° video, is a factor within the recording session. For a conventional, cinematic visual perspective, the visual impact of almost any immersive microphone array can be significantly reduced or eliminated through selective camera angles and editing, and the use of hanging microphones or discreet microphone stands. This has been the case since the very beginning of broadcast video. For immersive 360° visual environments, however, the visual impact of a large-scale microphone array could create a distracting cognitive disconnect for the user. One would assume that a compact, coincident, first-order or higher-order ambisonics-based microphone array would be best for this type of application, as they would have little to no visual impact. Depending on the context of the immersive visual and auditory environment, however, visible microphones might not be significantly distracting; the impact of visible microphones within virtual environments is a topic warranting further exploration.

5 Summary

A study was undertaken to compare three different immersive sound capture techniques, optimized for virtual rendering of acoustic music sound scenes, through binaural reproduction. The three techniques each represented a different principal in immersive microphone array design: spaced, near-coincident, and coincident. Stimuli were derived from 9-channel recordings of a solo piano in a highly reverberant environment. Results of a double-blind listening test showed:

(1) There were no significant differences between the spaced and near-coincident techniques for the perceptual attributes "envelopment," "naturalness of sound scene," and "naturalness of timbre." The spaced technique was shown to create a larger image of the sound source than the near-coincident technique.

(2) For all perceptual attributes under investigation, the coincident technique was rated significantly differently than the other techniques under investigation. This is similar to results seen in several previous studies.

Acknowledgements

This research was supported by the Social Sciences and Humanities Research Council and The Centre for Interdisciplinary Research in Music Media and Technology. Special thanks to Dr. Atsushi Marui.

References

[1] K. Hamasaki, K. Hiyama, "Development of a 22.2 Multichannel Sound System," *Broadcast Technology*, vol. 25 pp.9-13 (2006).

[2] T. Shibata, "Head mounted display," *Displays*, vol. 23, pp. 57-64 (2002).

[3] S. Kim, "Height Channels," In: A. Raginska, P. Geluso (ed) *Immersive Sound: The Art and Science of Binaural and Multichannel Audio*, New York: Routledge (2018).

[4] H. Lee, C. Gribben, "Effect of vertical microphone layer spacing for a 3D microphone array," *J Audio Eng Soc*, vol. 62, pp. 870–884 (2014).

[5] K. Hamasaki, W. Van Baelen, "Natural Sound Recording of an Orchestra with Three-Dimensional Sound," *Proc AES 141st Conv,* convention paper 9348 (2015).

[6] W. Howie, R. King, D. Martin, "A three-dimensional orchestral music recording technique, optimized for 22.2 multichannel sound," *Proc AES 141st Conv*, convention paper 9612 (2016).

[7] International Telecommunications Union. "Advanced sound system for programme production," *ITU-R BS.2051-2*, Geneva (2018)

[8] D. Griesinger, "Spatial Impression and Envelopment in Small Rooms," *Proc AES 103rd Conv*, convention paper 4638 (1997).

[9] K. Hamasaki, K. Hiyama, "Reproducing Spatial Impression With Multichannel Audio," *Proc AES 24th International Conference on Multichannel Audio*, conference paper 19 (2003).

[10] W. Howie, D. Martin, D. H. Benson, J. Kelly, R. King, "Subjective and objective evaluation of 9ch three-dimensional acoustic music recording techniques," *Proc AES Int Conf Spatial Reproduction — Aesthetics and Science*, conference paper P10-1 (2018).

[11] H. Lee, "The Relationship between Interchannel Time and Level Differences in Vertical Sound Localisation and Masking," *Proc AES 131st Conv*, convention paper 8556 (2011).

[12] R. Wallis, H. Lee, "The Reduction of Vertical Interchannel Crosstalk: The Analysis of Localisation Thresholds for Natural Sound Sources," *Applied Sciences*, vol. 7 (2017).

[13] E. Bates, S. Dooney, M. Gorzel, H. O'Dwyer, L. Ferguson, F. M. Boland, "Comparing Ambisonics Microphones–Part 2," *Proc AES 142nd Conv*, convention paper 9730 (2017).

[14] P. Geluso, "Capturing Height: The Addition of Z Microphones to Stereo and Surround Microphone Arrays," *Proc 132nd AES Conv*, convention paper 8595 (2012).

[15] A. Ryaboy, "Exploring 3D: A subjective evaluation of surround microphone arrays

catered for Auro-3D reproduction," *Proc AES 139th Conv*, convention paper 943 (2015).

[16] M. Gerzon, "Ambisonics in Multichannel Broadcasting and Video," *J Audio Eng Soc*, vol. 33, pp. 859-871 (1985).

[17] W. Howie, R. King, D. Martin, F. Grond, "Subjective evaluation of orchestral music recording techniques for three-dimensional audio," Proc *AES 142nd Conv*, convention paper 9797 (2017).

[18] T. Kamewaka, A. Marui, "Evaluation of recording techniques for three-dimensional audio recordings: Comparison of listening impressions based on difference between listening positions and three recording techniques," *Acoust Sci & Tech*, vol. 41, pp. 260-268 (2020).

[19] C. Millns, H. Lee, "An Investigation into Spatial Attributes of 360° Microphone Techniques for Virtual Reality," *Proc AES 144th Conv*, convention paper 10005 (2018).

[20] T. Kamekawa, A. Marui, A. Irimajiri, "Correspondence Relationship between Physical Factors and Psychological Impressions of Microphone Arrays for Orchestra Recording," *Proc AES 123rd Conv*, convention paper 7233 (2007).

[21] N. Peters, J. Braasch, S. McAdams, "Recording Techniques and their Effect on Sound Quality at Off-Center Listening Positions in 5.1 Surround Environments," *Canadian Acoustics*, vol. 41, pp. 37-49 (2013).

[22] A. Sitek, B. Kostek, "Study of Preference for Surround Microphone Techniques Used in the Recording of Choir and Instrumental Ensemble," *Archives of Acoustics*, vol. 36, pp. 365–378 (2011).

[23] G. Theile, H. Wittek, "Principals in Surround Recording with Height (v2.01)," *Proc 130th AES Conv*, convention paper 8403 (2011).

[24] EBU, "Loudness Metering: 'EBU Mode' Metering to Supplement EBU R 128 Loudness Normalization," *Tech 3341*, Geneva (2016).

[25] W. Howie, D. Martin, S. Kim, T. Kamekawa, R. King, "Effect of Audio Production Experience, Musical Training, and Age on Listener Performance in 3D Audio Evaluation," *J Audio Eng Soc*, vol. 67, pp.782–794 (2019).

[26] International Telecommunication Union, "Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems," *ITU-R BS.1116-3*, Geneva (2015).

[27] S. Kim, W. Howie, D. Martin, "Comparison of Salient Percepts Associated with Three Sound-Field Capturing Methods," *Proc ICSV25*, (2018).

[28] A. Gade, "Acoustics in Halls for Speech and Music," In: T. D. Rossing (ed) *Springer Handbook of Acoustics*, New York: Springer; pp.302-315 (2007).

[29] H. Lee, "Capturing 360◦ Audio Using an Equal Segment Microphone Array (ESMA)," *J Audio Eng Soc*, vol. 67, pp. 13–26 (2019).

[30] H. Wittek, G. Theile, "Development and application of a stereophonic multichannel recording technique for 3D Audio and VR," *Proc AES 143rd Conv*, convention paper 9869 (2017).

[31] H. Lee, "Multichannel 3D Microphone Arrays: A Review," *J Audio Eng Soc*, vol. 69, pp. 5-26 (2021).