Audio Engineering Society

Conference Paper 9

Presented at the International Conference on Spatial & Immersive Audio,

2023 August 23-25, Huddersfield, UK

Objective comparisons of 3D audio reproduction with and without bottom channels

Will Howie[1], Atsushi Marui[2], Toru Kamekawa[2], Florian Grond[3], and Akira Omoto[4]

[1]Japan Society for the Promotion of Science International Research Fellow, Tokyo University of the Arts, Tokyo, Japan
[2]Department of Musical Creativity and the Environment, Tokyo University of the Arts, Tokyo, Japan
[3]Concordia University, Montreal, Canada
[4]Faculty of Design, Kyushu University, Fukuoka, Japan

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

ABSTRACT

This study examines several different methods for measuring objective differences between 3D audio reproduction conditions with and without bottom channels (floor-level loudspeaker channels) within 9+10+8 audio reproduction. The methods, derived from previous studies investigating 2D and 3D audio reproduction, examine mono and binaural signal features, as well as various ratios of directional sound energy. Stimuli were created using short excerpts of seven different 3D audio recordings covering a range of musical and non-musical sound scenes and audio recording methods. Three different reproduction conditions were examined: 1) all vertical loudspeaker layers active, 2) bottom-layer signals muted), and 3) downmix: bottom-layer signals and main-layer signals merged, across three different acoustic environments: a large mixing studio, a small mixing studio, and a hemi-anechoic room. In this study, most measurement techniques under test did not yield meaningful results. However, averaged power spectra measurements show a consistent trend towards greater low frequency energy when the bottom-layer loudspeaker channels are active. Possible alternative measurement techniques are discussed.

1 Introduction

This paper examines the implementation of several different sound field measurement and analysis techniques designed or adapted to quantify possible objective effects of bottom channels, i.e., lower-elevated loudspeakers, in 3D audio reproduction. Numerous 3D audio reproduction formats have been introduced for both loudspeaker and binaural headphone-based mediums, many of which have been standardized or summarized by the International Telecommunications Union (ITU) [1]. There is general agreement among researchers that for a wide range of 3D audio content, the inclusion of sonic information from elevated "height channels" increases listener impression of perceptual factors such as depth, presence, envelopment, naturalness, realism, and intensity [2–5]. However, comparatively little research has investigated the effect of lower-elevated or floor-level loudspeakers. Whereas height channels are often associated with ambient information, especially for acoustic music recording and mixing [6], the bottom channels tend to be used more for panning of direct sounds, be it for matching on-screen action [7], or extending vertical sound images to the floor, to better match real life listening experiences [8]. As such, we would expect the bottom channels to affect the listening experience in different ways than height channels.

A study investigating the perceptual effects of 3D audio reproduction with bottom channels is underway at Tokyo University of the Arts, McGill University, and Rochester Institute of Technology, with results forthcoming. That study focuses on uncovering subjective differences between three different reproduction conditions (complete, no bottom layer, and downmix) of seven different 3D sound scenes featuring bottom-layer sonic information.

A subset of the same researchers wished to investigate the possibility of using physical measurements to also quantify differences between those three reproduction conditions; at present, the authors are not aware of any previous study that has specifically examined the physical influence of lower-elevated direct or reflected sound on 3D audio reproduction. The goals of the current study are:

1) Examine the appropriateness of several sound field measurement techniques in terms of quantifying 3D audio reproduction with and without bottom channels.
2) Collect objective data that can eventually be compared with subjective perceptual data from highly experienced listeners.

2 Background

2.1 3D audio reproduction with bottom channels

Several commercially available immersive audio formats already include bottom-layer loudspeaker-based sound reproduction, such as NHK 22.2 Multichannel Sound [7] and Sony 360 Reality Audio [9]. Various prototype or experimental 3D audio systems also include bottom-layer sound reproduction, such as Yamaha's "ViReal" [10], the "Sound Cask" (Kyushu University and Tokyo Denki University) [11], and Hitachi's "Tesseral Array" [12], while many binaural software renderers and plugins include the option for negative vertical panning. Ambisonics-based microphones, a popular method of sound capture for 360 video and virtual and augmented reality platforms, capture sound both above and below a vertical centre-reference point, and so should be a natural fit for systems that support reproduction of sound from below the listener.

2.2 9+10+8 audio reproduction

The 3D audio stimuli used in this study, and above-mentioned perceptual study, were recorded or rendered specifically for a 9+10+8 reproduction system. Following the ITU's naming convention for advanced audio systems [1], 9+10+8 refers to nine height channels, ten "main layer" channels (loudspeakers placed roughly at ear-level), and eight bottom channels. 9+10+8 is identical to 9+10+3 (NHK 22.2) in terms of number and spatial positions of loudspeakers, but adds bottom channels for the Side Left and Right, and Rear Left, Centre, and Right speaker positions (Fig. 1). The addition of the five bottom channels gives an even spatial distribution of loudspeakers in all three vertical layers.

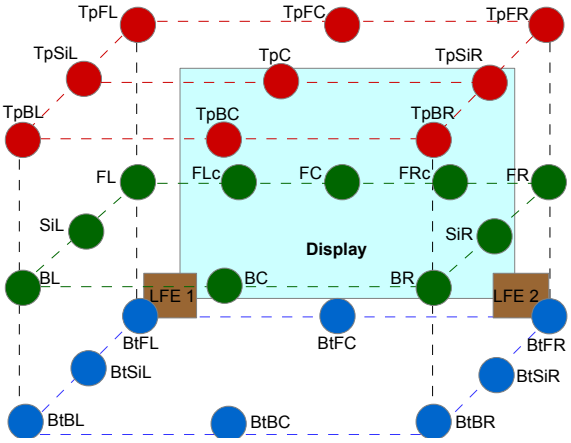

Figure 1. 9+10+8 channel/loudspeaker layout. Channel naming convention as per ITU [1].

2.3 Objective measures of 3D sound scenes

Various objective measurements have been developed to quantify the acoustic properties of a given space [13]. These measurements normally examine captured mono or binaural room impulse responses, and do not necessarily translate to measuring playback of continuous multichannel audio signals, such as the type used in this study. However, several previous studies have shown evidence of successfully adapting room acoustic measurement techniques for comparing multichannel audio signals [14–18].

In a study comparing different 3D acoustic music recording techniques, Howie et al. [17] found measured Interaural Cross-Correlation Coefficient (IACC) values to be a good predictor of listener perception of the subjective attributes "sound source image size" and "envelopment". In that same study, Schuitman et al.'s "P_{ASW}" model, which also measures binaural signals, was found to be predictive of "sound source image size", while the mono signal feature "Spectral Variation" was found to be predictive of "envelopment". Kamekawa and Marui [18] found some correlation between subjective listener evaluations of various 3D music recording techniques, and the objective measurements "Spectral

Centroid," "Lateral Reflected Energy," and "Vertical Reflected Energy."

Power et al. found a strong negative correlation between mean listener "envelopment" scores for various 2D and 3D multichannel reproduction systems under investigation and measured IACC values for binaural dummy-head recordings made of the testing stimuli [14]. Similarly, Choisel and Wickelmaier [15] reported a strong negative correlation between $IACC_f$ (a variation of IACC that includes a pre-processing step designed to mimic envelope extraction in the human auditory system) values and perceived "spaciousness" when comparing $IACC_f$ measurements of binaural recordings of multichannel audio stimuli with listener evaluations. Masson and Rumsey [16] found perceptually grouped IACC measurements on experimental stimuli correlated highly with listener subjective data.

3 Method

3.1 Creation of 3D audio stimuli

Short excerpts, 20-30 s each, of seven 3D sound scenes were selected as stimuli for this study, as well as the concurrent subjective study. The sound scenes were selected in an attempt to cover a range of musical and non-musical content and 3D audio production styles within a compact set of stimuli. Brief descriptions of each sound scene and their respective production methodologies follow below. More documentation, and methodological and aesthetic explanations of the music recordings can be found in [8] and this online audio/visual repository: https://doi.org/10.5281/zenodo.7563813

Rock Music

An alternative rock song was recorded in a large recording studio (floor area = 160 m^2, ceiling height = 7 m, reverb time = approx. 1.0 s at 500 Hz.) The song features a dense musical arrangement with a central lead vocal, and various acoustic and electronic instruments spread around the listener in horizontal and vertical space, combining "realistic" and "hyper-realistic" aesthetics. Instruments were generally captured with a combination of complex, multichannel-close microphone arrays and widely spaced ambience microphones. For the "cut" version of this sound scene, certain direct sound microphone signals that were panned to the bottom layer for aesthetic reasons in the "Full" version, such as Kick Drum or Bass Amp, were retained and panned to the main layer in order to maintain continuity within the musical arrangement and balance.

Solo Piano

A solo piano performance of a jazz standard, recorded in the same large recording studio as above. An array of nine cardioid microphones, arranged as Left, Centre, and Right in three vertical layers, was placed near the piano to capture primarily direct sound, panned to create an image of the concert grand piano that gives the listener a realistic impression of the instrument's horizontal and vertical extent. An array of largely spaced directional microphones captured ambient sound for the height-layer speakers, and side and rear main-layer and bottom-layer speakers.

Pipe Organ

A performance of solo pipe organ music was recorded at Tokyo University of the Arts' Sōgakudō Concert Hall (length = 36 m, width = 18 m, height = 15 m, capacity = 1100 seats, reverb time = 2.4 s). This is an example of a microphone setup optimized for a stereo recording (Decca tree with outriggers, room mics, and "close" mics) being augmented with additional microphones to create a 3D recording.

Solo Bass (hemi-anechoic environment)

A double bass improvising in a contemporary jazz style was recorded in a hemi-anechoic environment to include a presentation of an acoustic instrument free from interactions between direct and reflected sound. The recording venue was the Spatial Audio Laboratory at the Centre for Interdisciplinary Research in Music Media and Technology (reverb time = 0.1 s). A simple 5 microphone array (Left, Centre, Right, Top, Bottom) captured the direct sound of the instrument.

Waterfall

This outdoor recording was made using a quasi-spherical near-spaced microphone array of 24 DPA 4017c shotgun microphones, developed by Omoto and Kashiwazaki at Kyushu University. Details on the microphone array's design philosophy and construction, which are based on the Boundary Surface Control principal, can be found in [19] and [20]. The microphones are distributed evenly at every 45° in elevation and azimuth angles: eight microphones per layer, for three vertical layers. These microphone signals correspond directly to their respective loudspeakers (e.g., the main layer 45° Left signal would be routed to the FL channel). No signals were routed to the FLc, FRc or TpC channels. The recording was made near a waterfall in the town of Takachiko, Japan, with the array positioned to capture the sound of "water trickling from rocks," with the

waterfall imaging in front of the listener. This resulted in a densely rich 360° sound scene of cascading water, with an even dynamic envelope. The recording was made by members of the Department of Acoustic Design at Kyushu University's OMT Lab: http://www.design.kyushu-u.ac.jp/~omotoken/

Urban Soundscape
This recording of outdoor ambience was made at Roy Terrace Community Gardens in Montreal, Canada, by a researcher highly experienced in ambisonics recording techniques. The sound scene was captured using an em32 Eigenmike from MH Acoustics, a spherical coincident microphone array of 32 pressure capsules. The em32 was placed in roughly the centre of this open plaza/garden space, at a height of approximately 1.5 m, in an attempt to capture a complete and equal sonic perspective of area. Microphone signals were converted into a 4th-order ambisonics B-format audio file using the provided "EigenStudio" software, following ACN channel order with N3D normalization [21]. An excerpt including the sounds of a passing cyclist, a skateboarder, and a child playing was decoded for 24channel reproduction (9+10+8, omitting the FLc, FRc, and TpC loudspeakers) using an open-source AllRAD [22] decoder plugin from IEM.

Taiko Drum Ensemble
A recording was made of a taiko drum ensemble (one ōdaiko and two shime-daiko) in the above-mentioned large studio. A simple, largely-spaced one microphone per loudspeaker setup was used, with primarily directional microphones for both direct and diffuse sound capture.

All music recordings were made at 96 kHz/24 bit resolution to either Avid Pro Tools or Merging Technologies Pyramix workstations by a team of audio engineers and researchers with extensive experience recording and mixing 3D audio. Outdoor ambience recordings were captured at 48 kHz/24 bit resolution, then converted to 96 kHz resolution to facilitate playback together with the other recordings. Recordings were mixed or rendered at Tokyo University of the Arts Senju Campus' Studio B (see: section 3.2). The two non-musical sound scenes were selected from a number of available indoor and outdoor recordings, based on their effective capture and presentation of sound coming from below the listener. Three versions of each of the seven sound scene excerpts were created, for a total of 21 stimuli:

- "Full": the original 9+10+8 mix or rendering
- "Cut": all bottom channel signals from the 9+10+8 mix removed
- "X": a downmix wherein the bottom channel signals have been mixed with their corresponding main-layer signals (e.g., BtFC + FC = FCx) at a 1:1 ratio, with the new signals being reproduced only from the main layer loudspeakers.

For each sound scene, playback of all stimuli was level-matched to within 0.1LUFS of each other by means of a B&K Type 4128 Head and Torso Simulator situated at the listening position, and a software loudness meter (integrated measurements, EBU +9 scale).

3.2 Reproduction environments

Measurements of stimuli playback were taken in three different rooms equipped with 3D multichannel audio reproduction systems. The rooms were chosen based on availability of spaces with compatible loudspeaker layouts, and to cover a range of acoustic environments.

Studio B, Tokyo University of the Arts
This large room (floor area = 68 m^2, ceiling height = 5 m) is located at Tokyo University of the Arts' Senju Campus. Originally built as a recording space, the room has an acoustically treated ceiling and walls, and a reflective hardwood floor. Though the room's reverb time is somewhat longer than is typical of critical listening environments (ca. 0.4 s at 500 Hz), the space otherwise conforms to ITU BS.1116 [23] recommendations for critical listening. Studio B is equipped with 27 KS Digital C5 2-way powered studio monitors, which have a fairly linear frequency range from 48 Hz to 22 kHz. Speaker positions conform to ITU recommendations for 9+10+3 reproduction [1], with the five added bottom channels matching the horizontal angles of their corresponding main-layer speakers (Fig. 2).

Studio SP, Tokyo University of the Arts
Studio SP, also located at Tokyo University of the Art's Senju Campus, is a multichannel mixing studio whose size and acoustics are contrasting to those of Studio B: (Floor surface area = 41.3 m^2, ceiling height = 2.4 m, reverb time = 0.2 s at 500 Hz). Acoustically, this space more resembles a broadcast studio, cinema, or even home theatre environment, with acoustically treated walls and ceiling, and a carpeted floor. SP is equipped with 27 Revox Piccolo s60 passive 2-way monitors (frequency response: 65 Hz – 20 kHz),

powered by an Innosonix MA 32 amplifier. Speaker positions are identical to those of Studio B in terms of azimuth and angles of elevation.

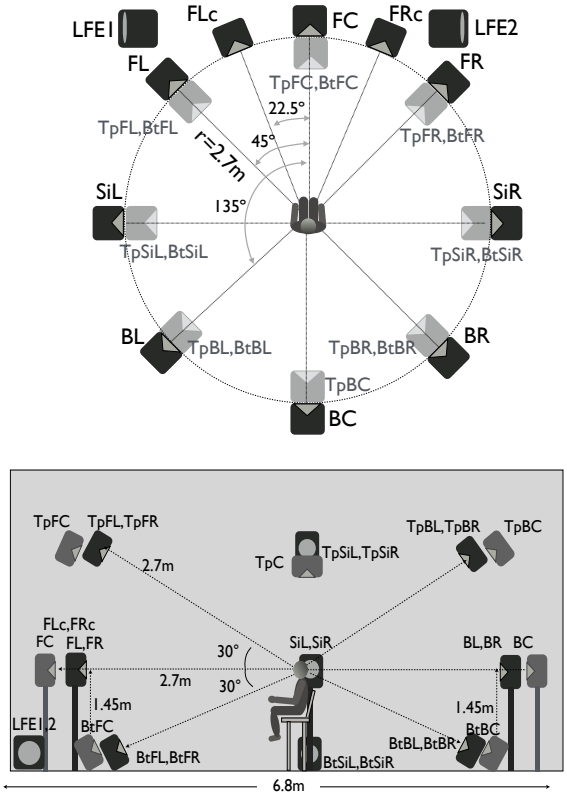

Figure 2. Studio B speaker layout, overhead and side views.

Hemi-anechoic room, Kyushu University

The Hemi-anechoic room is located at Kyushu University, Ohashi Campus, and is purpose-build for acoustics and multichannel audio research: (W = 7.1 m, L = 6.4 m, H = 2.5 m, RT60 < 0.1 s, background noise below NC-15). For this experiment, additional absorptive material was added to the floor in a 1.5 m radius around the location of the measurement microphones to dampen floor reflections. The room is equipped with 24 Genelec 8020 loudspeakers (frequency response: 62 Hz - 20 kHz ± 2.5 dB) setup in three identical vertical layers. The horizontal speakers are all spaced 45° apart from neighbouring speakers, with a vertical displacement of 30° between each speaker layer. Two Genelec 8040 loudspeakers were added to the main layer at ±22.5° to match the physical layout of the speaker positions within the other two reproduction rooms. The Hemi-anechoic space did not include a TpC channel: an acceptable compromise, as none of the sound scenes under investigation had significant audio content mixed to that channel.

3.4 In-room measurements

For each acoustic environment, playback of the stimuli under investigation was captured using the following microphone setups, as per implementation of these measurements in [17], [18], and [24]:

- Front facing omnidirectional (DPA4006) and a coincident, laterally-oriented bi-directional microphone (Schoeps MK8)
- Front facing omnidirectional (DPA4006) and a coincident, vertically-oriented bi-directional microphone (Schoeps MK8)
- Front facing bi-directional, and coincident, vertical bi-directional microphones (Schoeps MK8)
- Head and Torso Simulator
    - B&K Type 4128 (Studios B/SP)
    - HeadAcoustics HSU (Kyushu)

Microphones were placed at a point equidistant to all main-layer loudspeakers, at the same height as the centre-point of the FC loudspeaker. Microphone signals were routed to an RME Fireface UFX for pre-amplification and analog-to-digital conversion, and recorded at 96 kHz/24 bit resolution.

3.5 Selected measurement techniques

Based on a review of results from previous studies including objective measurements of multichannel audio sound scenes discussed in Section 3.3, several sound field capture and analysis techniques were selected for test in the current study, and are summarized below. Though IACC features prominently within the review, it is typically within the context of investigating correlation with subjective results, which is not within the scope of the current study. Also, it has been shown that the just noticeable difference (JND) for IACC tends to cover a wide range of values depending on the program material used [25–28]. As such, IACC will not be examined as an objective measurement technique in this study, though the authors did make appropriate recordings using Head and Torso Simulator (HATS) microphones at each venue (Section 3.4) for possible future comparisons of IACC values with subjective listener data.

Binaural features

A set of signal features were derived from Schuitman et al.'s [29] binaural model designed to predict room acoustic attributes. These features, named $P_{REV,}$

P_{CLAR}, P_{ASW}, and P_{LEV}, have been shown to correlate with subjective assessments of "reverberance", "clarity", "apparent source width" and "listener envelopment", respectively. This set of measurements was implemented using the stereo signals from the binaural HATS recordings of the stimuli. It was hoped that these measurements could be used to quantify differences in spatial impression between the reproduction condition conditions under test, a similar motive to the authors of [17].

Monaural spectral features
Using a mono summation of the HATS signals and the open-source Timbre Toolbox's [30] "ERBfft" setting, the following monaural spectral features were calculated: Spectral Centroid, Spectral Crest factor, Spectral Flatness, Spectral Kurtosis, Spectral Skew, Spectral Spread, and Spectral Variation. Additionally, the averaged power-spectra of the playback of each stimulus were derived from an omnidirectional microphone signal. These measurements were all aimed at quantifying timbral differences between reproduction conditions.

Sound Energy Ratios
The following objective features were calculated to investigate various physical sound energy ratios:
1) Ratio of Front to Side Energy (FSER): omni and lateral oriented bi-directional microphones.
2) Ratio of Front to Vertical Energy (FVER): omni and vertical oriented bi-directional microphones.
3) Ratio of Front to Overhead Energy (FOER): a variation of (2) devised by Kim [24], using front facing and vertically oriented bi-directional microphones.

FVER and FOER were chosen as they are specifically designed to compare the amount of vertical sound energy present in multichannel reproduction with total frontal sound, though for a 360° sound scene both of their associated microphone techniques would also capture rear sound in the horizontally oriented microphone. With these techniques, the goal was to discover if distinct differences in vertical imaging and panning between reproductions conditions could be measured objectively. Since loudspeakers are positioned at ±90° in all three vertical layers of the 9+10+8 format, FSER was also included to see if any quantifiable changes in side energy occurred between reproduction conditions.

4 Results

4.1 Sound energy ratio measurements

No systematic trend was observed within the measurements for FSER, FVER, or FOER in terms of differences between the three reproduction conditions (Full, X, Cut). For specific sound scenes, some differences between reproduction conditions were observed within specific reproduction rooms, but these differences were not consistent across all three rooms.

4.2 Monaural and binaural features

For the monaural spectral features calculated using the Timbre Toolbox, again, no systematic trend was observed within the output of values. The same was true for the values of P_{REV}, P_{CLAR}, P_{ASW}, and P_{LEV}.

4.3 Power spectrum measurements

An examination of the averaged power spectra of the stimuli under investigation found a clear trend wherein the Full condition displays a greater amount of low frequency energy than the X and Cut conditions. This trend was found across all three rooms, and for all sound scenes except "Urban Soundscape". Figures 3–5 show the combined mean spectral responses of all sound scenes, excluding "Urban Soundscape," for each reproduction condition, for each room. As can be seen, the difference in low frequency energy varies between sound scenes, ranging from approximately 2–10 dB at the center frequency, which itself varies between reproduction rooms.

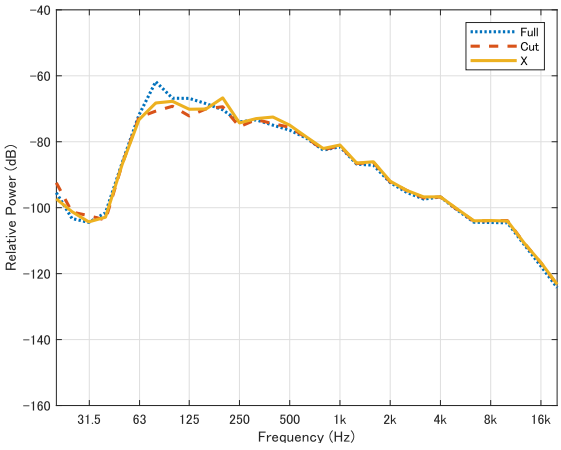

Figure 3. Combined mean power spectra for all sound scenes, Studio B.

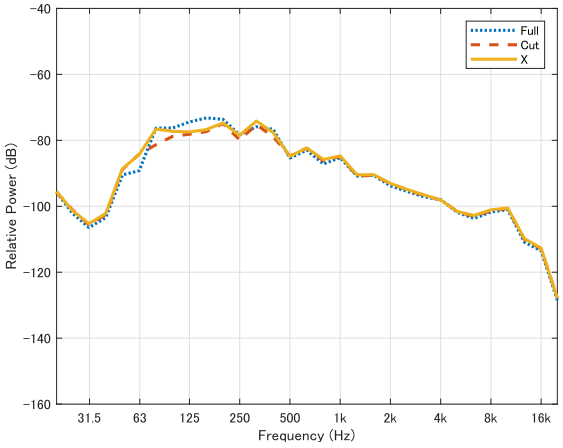

Figure 4. Combined mean power spectra for all sound scenes, Studio SP.

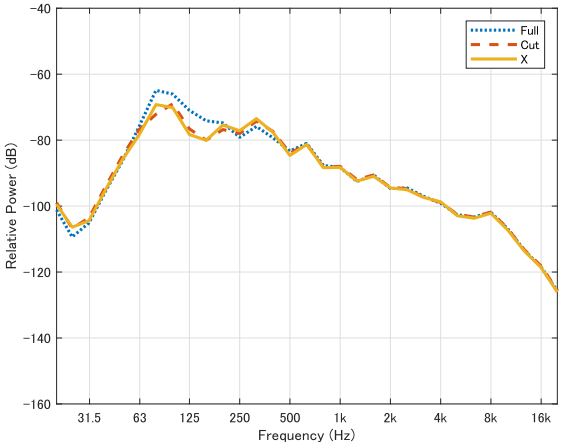

Figure 5. Combined mean power spectra for all sound scenes, Hemi-anechoic room.

5 Discussion

5.1 Variation in measured power spectra

As seen in Section 4, within the objective features under investigation in this study, the most consistent difference between reproduction conditions with and without bottom channels appears to be an increase in low frequency energy observed within the Full condition, as compared with both the Cut and X conditions. This is not so surprising for the Full vs Cut comparisons: we would assume to find noticeable spectral differences between playback conditions containing different combinations of microphone signals. Several of the musical instruments featured within the sound scenes, such as the piano or double bass, naturally radiate a darker tonality and more lower frequency content from areas physically lower or underneath [31], and this information may have been best captured by the bottom reproduction layer's corresponding microphones.

Cabrera and Tilley state that "on a reflective floor for a normal range of ear heights (from seated to standing), the spectral notches caused by interference between direct and reflected sound extend to low frequencies for elevated sources – meaning that a source on or near the floor will tend to convey more bass than an equivalent elevated source. [32]" This theoretical increase of low frequency transmission efficiency for floor-level loudspeakers may explain why for most of the sound scenes, across all three reproduction rooms, we can observe more low frequency information in the Full condition as compared with the X condition. Recall that both conditions contain the same balance of identical microphone signals: the only difference being that in the X condition, the main and bottom layer signals are both being reproduced from the main layer loudspeakers. The close proximity of the bottom channel loudspeakers to the floor, a reflective surface for low frequencies across all three venues, may also be a factor in boosting the frequency ranges seen in Figures 3–5.

It is interesting to consider the above within the context of the original purpose of including bottom channels in 3D audio reproduction: achieving greater realism and flexibility within vertical panning [7]. The general increase in low-frequency transmission is likely an unintended, but potentially very useful outcome for content creators. Audio engineers working with bottom-layer loudspeakers should be aware of the potential loss of low frequency information when downmixing to other formats.

5.2 FSER, FVER, FOER

In this study, the physical measurements of FSER, FVER, and FOER were not found to be useful in terms of measurably quantifying or defining differences between the various reproduction conditions under investigation. This result is rather surprising for the FVER and FOER measurements, given that the Full condition certainly outputs more physical sound energy from below than the other two conditions. There are several possible explanations, which are likely correlated to a certain degree. Firstly, both these measures are designed to look at total vertical energy (positive and negative lobes of the figure-8 microphone). It may be that even when the bottom layer signals were inactive, the level of top-layer signals was sufficient for the analysis to show similar levels of sound energy between reproduction conditions. Another factor to consider is floor reflections: the floor in Studio B is highly reflective,

and while carpeting or acoustical treatment covered some or all of the floor surface in Studio SP and the Hemi-Anechoic space, low frequency reflections from main-layer loudspeaker content would still have been present. These reflections, if captured by the downward facing lobe of the bi-directional microphone, would likely skew measurement results. Lastly, the choice of a vertically oriented bi-directional microphone may itself have been a problem. The negative angle of elevation of the floor-level loudspeakers in each room is -30°. As such, the transmission path from the loudspeakers to the microphone may have resulted in a certain degree of sound energy being rejected by the null point in the microphone's polar pattern.

A better solution may be to compare ratios of sound energy captured with an omnidirectional microphone (total energy) versus a dual-diaphragm condenser microphone where the output signals of both diaphragms are accessible (Sennheiser MHK800 Twin, Austrian Audio OC818, etc.). One could then compare the omnidirectional signal with either the downward or upward facing cardioid signals, depending on what direction of vertical sound energy is being investigated. Similarly, if the sound field was captured with an ambisonics-based microphone, any number of virtual microphone signals of specific direction and degree of directivity could be generated for a more focused comparison against the omnidirectional signal. This would also allow for visual comparisons of sound intensity distribution between different stimuli [33].

5.3 Other measures under test

In previous studies comparing microphone techniques for 3D audio reproduction, Schuitman et al.'s P_{ASW} model was found to be a good predictor of differences in spatial impression between stimuli [17], while the timbre-related measurements Spectral Variance and Spectral Centroid were also found to correlate somewhat with listener impressions [17, 18]. Different microphone techniques are designed with specific goals in terms of spatial impression and timbral fidelity, and so we would expect to see measurable differences between their output. In the current study, however, any differences in spatial impression or timbre between the reproduction conditions under test, beyond those observed in the low frequencies, appear to be too subtle to generate meaningful data with these measurement methods. By combining spectral and spatial features, it may be possible to predict perceptual impressions of differences between these reproduction conditions.

Acknowledgements

This work was supported by the Japan Society for the Promotion of Science, Tokyo University of the Arts, and JSPS KAKENHI grant numbers 21H03764 and 21F21745.

References

[1] International Telecommunications Union, "Advanced sound system for programme production," (2022).

[2] J. Francombe, T. Brookes, R. Mason, and J. Woodcock, "Evaluation of Spatial Audio Reproduction Methods (Part 2): Analysis of Listener Preference," *J. Audio Eng. Soc.*, vol. 65, no. 3, pp. 212–225 (2017).

[3] S. Oode, I. Sawaya, K. Ono, and K. Ozawa, "Three-Dimensional Loudspeaker Arrangement for Creating Sound Envelopment," *IEICE Tech. Rep.*, vol. 112, no. 125, pp. 7–12 (2012).

[4] T. Kamekawa, A. Marui, T. Date, and M. Enatsu, "Evaluation of spatial impression comparing 2ch stereo, 5ch surround, and 7ch surround with height channels for 3D imagery," presented at the *130th Convention of the Audio Engineering Society* (2011).

[5] S. Kim, D. Ko, A. Nagendra, and W. Woszczyk, "Subjective Evaluation of Multichannel Sound with Surround-Height Channels," presented at the *135th Convention of the Audio Engineering Society* (2013).

[6] Lee, "Multichannel 3D Microphone Arrays: A Review," *J. Audio Eng. Soc.*, vol. 69, no. 1/2, pp. 5–26 (2021).

[7] K. Hamasaki and K. Hiyama, "Development of a 22.2 Multichannel Sound System," *Broadcast Technology*, vol. 25, no. Winter, pp. 9–13 (2006).

[8] W. Howie, T. Kamekawa, M. Morinaga, "Case Studies in Music Production for Advanced 3D Audio Reproduction with Bottom Channels," presented at the *AES International Conference on Spatial & Immersive Audio* (2023).

[9] A. Nakai, M. Tsuji, and T. Chinen, "Directional Dependency of Subjective Sound Pressure Perception on Three-Dimensional Sound," presented at the *148th Convention of the Audio Engineering Society* (2020).

[10] S. Kaneko et al., "Development of a 64-Channel spherical microphone array and a 122-channel loudspeaker array system for 3D sound field capture and reproduction technology research," presented at the *144th Convention of the Audio Engineering Society* (2018).

[11] A. Omoto et al., "Sound field reproduction and sharing system based on the boundary surface control principle," *Acoust. Sci. Tech.*, vol. 36, no. 1, pp. 1–11 (2015).

[12] Y. Tanabe, G. Yamauchi, A. Marui, and T. Kamekawa, "Tesseral Array for Group-Based Spatial Audio Capture and Synthesis," presented at the *International Conference on Audio for Virtual and Augmented Reality* (2020).

[13] T. D. Rossing, "Acoustics in Halls for Speech and Music," *Springer Handbook of Acoustics*, Springer (2007).

[14] P. Power et al., "Investigation into the Impact of 3D Surround Systems on Envelopment," presented at the *137th Convention of the Audio Engineering Society* (2014).

[15] S. Choisel and F. Wickelmaier, "Relating auditory attributes of multichannel sound to preference and to physical parameters," presented at the *120th Convention of the Audio Engineering Society* (2006).

[16] R. Masson and F. Rumsey, "A comparison of objective measurements for predicting selected subjective spatial attributes," presented at the *112th Convention of the Audio Engineering Society* (2002).

[17] W. Howie et al., "Subjective and objective evaluation of 9ch three-dimensional acoustic music recording techniques," presented at the *AES International Conference on Spatial Reproduction – Aesthetics and Science* (2018).

[18] T. Kamekawa and 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, no. 1, pp. 260–268 (2020).

[19] A. Omoto and H. Kashiwazaki, "Hypotheses for constructing a precise, straightforward, robust and versatile sound field reproduction system," *Acoust. Sci. Tech.*, vol. 41, no. 1, pp. 151–159 (2020).

[20] H. Kashiwazaki and A. Omoto, "Sound field reproduction system using narrow directivity microphones and boundary surface control principle," *Acoust. Sci. Tech.*, vol. 39, no. 4, pp. 295–304 (2018).

[21] M. Chapman, W. Ritsch, T. Musil, I. Zmo ̈lnig, H. Pomberger, F. Zotter, *et al.*, "A Standard For Interchange of Ambisonic Signal Sets: Including a file standard with metadata," presented at the *Ambisonics Symposium 2009* (2009).

[22] F. Zotter and M. Frank, "All-Round Ambisonic Panning and Decoding," *J. Audio Eng. Soc*, vol. 60, no. 10, pp 807–820 (2012).

[23] International Telecommunications Union, "Methods for the Subjective Assessment of Small Impairments in Audio Systems," (2015).

[24] S. Kim et al., "Height loudspeaker position and its influence on listeners' hedonic responses," presented at the *AES Conference on Sound Field Control* (2016 July).

[25] A. Walther and C. Faller, "Interaural correlation discrimination from diffuse field reference correlations," *J. Acoust. Soc. Am.* vol. 133, no. 3, pp. 1496–1502 (2013).

[26] K. Fujii et al., "Spatial Distribution of Acoustical Parameters in Concert Halls: Comparison of Different Scattered Reflections," *Journal of Temporal Design in Architecture and the Environment* vol. 4, no. 1 pp. 59–68(2004)

[27] T. Okano, "Judgments of noticeable differences in sound fields of concert halls caused by intensity variations in early reflections," *J. Acoust. Soc. Am.* vol. 111, no. 1 pp. 217–229 (2002).

[28] J. S. Bradley, "Review of objective room acoustics measures and future needs," in *Proceedings of the International Symposium on Room Acoustics* (2010).

[29] J. van Dorp Schuitman and D. de Vries, "Deriving content-specific measures of room acoustic perception using a binaural, nonlinear auditory model," *J. Acoust. Soc. Am.* vol. 133, no. 3, pp. 1572-1585 (2013).

[30] G. Peeters et al., "The Timbre Toolbox: Extracting audio descriptors from musical signals," *J. Acoust. Soc. Am.*, vol. 130, no. 5, pp. 2902-2916 (2011).

[31] J. Meyer, *Acoustics and the Performance of Music*, Springer (2009).

[32] D. Cabrera and S. Tilley, "Vertical Localization and Image Size Effects in Loudspeaker Reproduction," presented at the *AES 24th International Conference on Multichannel Audio*, (2003).

[33] M. Nakahara, A. Omoto, and Y. Nagatomo, "A simple evaluating method of a reproduced sound field by a measurement of sound intensities using Virtual Source Visualizer," presented at the *140th Convention of the Audio Engineering Society* (2017 Oct).