OPEN ACCESS Freely available online

W. Howie, D. Martin, A. Marui, T. Kamekawa,
S. Kim, A. Aydin, and R. King, "Perceptual Comparison of 3D Audio
Reproduction With and Without Bottom Channels,"
J. Audio Eng. Soc., vol. 72, no. 4, pp. 221–234 (2024 Apr.).
https://doi.org/10.17743/jaes.2022.0136.

Perceptual Comparison of 3D Audio Reproduction With and Without Bottom Channels

WILL HOWIE,[1]* *AES Associate Member*, **DENIS MARTIN,**[2] *AES Associate Member*,
(wghowie@gmail.com) (denis.m.martin@gmail.com)
ATSUSHI MARUI,[3] *AES Member*, **TORU KAMEKAWA,**[3] *AES Member*, **SUNGYOUNG KIM,**[4,5] *AES Member*,
(marui@ms.geidai.ac.jp) (kamekawa@ms.geidai.ac.jp) (sxkiee@rit.edu)
AYBAR AYDIN,[6] *AES Student Member* **AND RICHARD KING,**[6] *AES Fellow*
(aybar.aydin@mail.mcgill.ca) (richard.king@mcgill.ca)

[1] *Japan Society for the Promotion of Science International Research Fellow, Tokyo University of the Arts, Tokyo, Japan*
[2] *Faculty of Music, University of Toronto, Toronto, Canada*
[3] *Department of Musical Creativity and the Environment, Tokyo University of the Arts, Tokyo, Japan*
[4] *Graduate School of Culture Technology, Korea Advanced Institute of Science and Technology, Daejeon, Korea*
[5] *College of Engineering Technology, Rochester Institute of Technology, Rochester, USA*
[6] *Graduate Program in Sound Recording, McGill University, Montreal, Canada*

This study examines the perceptual effects of bottom channels, i.e., floor-level loudspeakers, within 3D audio reproduction. Two listening tests were undertaken at three different venues, using experienced subjects. Both experiments involved comparing three different versions of seven different musical and nonmusical sound scenes: the original mix with all three vertical loudspeaker layers active (Full), the bottom layer muted (Cut), and the bottom layer downmixed into the main layer loudspeakers (X). Results indicate that listeners could discriminate between the three reproduction conditions with a very high degree of accuracy, particularly when comparing the "Full vs. Cut" and "Full vs. X" conditions. Subjects found that the most salient aspects of the sound scene in terms of differentiating between reproduction conditions were related to low-frequency energy, changes in horizontal and vertical imaging, and timbre/tone. Discrimination ability between reproduction conditions was consistent across all three listener groups, though subjects' perception of the degree of difference between reproduction conditions across various auditory attributes varied between groups. These differences may be related to subjects' previous experience with 3D audio including bottom channels, venue bottom-layer loudspeaker angles of elevation, and venue acoustic conditions.

0 INTRODUCTION

A study has been undertaken at Tokyo University of the Arts to investigate the effect of bottom channels (i.e., lower-elevated or floor-level loudspeakers) in 3D audio reproduction. This paper focuses on perceptual differences observed within subjective comparisons of three different reproduction conditions: full mix with all vertical loudspeaker layers, no bottom layer, and a downmix merging the bottom and main layers. The stimuli, which cover a range of musical and nonmusical sound scenes, were created and rendered for a 9+10+8 (27.2) 3D audio reproduction environment. A semiconcurrent study examining objective measurement techniques for quantifying differences between these three reproduction conditions is described in a separate paper [1] and summarized in SEC. 1.1.

Numerous commercial and broadcast 3D audio formats have been introduced, many of which have been summarized or standardized by the International Telecommunications Union (ITU) [2]. These audio formats aim to provide listeners with a spatial impression that augments a sense of reality, ambience, and envelopment while maintaining excellent sound quality and sound image stability across a wide viewing and listening area [3]. Typically, 3D sound fields are reproduced using either an array of loudspeakers or over headphones using binaural rendering [4].

Several commercially available and prototype loudspeaker-based 3D audio systems already include the capacity to position sound below the listener [2, 5–9],

*Correspondence should be addressed to: Will Howie, e-mail: wghowie@gmail.com, Last updated: March 13, 2024.

while many binaural audio rendering tools also have full vertical panning control. Previous research has shown that for a wide range of immersive 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 [10–14]. Comparatively little research, however, has examined the effect of bottom channels. This study, therefore, aims to contribute to a more complete understanding of the perceptual influence of lower-elevated sound in 3D audio reproduction.

1 BACKGROUND

1.1 Previous Research

Numerous previous studies have examined the perceptual effects of height channels on audio reproduction, many of which are effectively summarized by Roginska and Geluso [4] and Paterson and Lee [15]. At this time, however, only two studies are known to have specifically examined the perceptual effects of bottom channels in 3D audio reproduction.

In a study by Grewe et al. [16], subjects evaluated audio stimuli that were either recorded specifically for or remixed for a 4+5+3 reproduction system (i.e., four height channels, four main layer channels, and three bottom channels, as per ITU-recommended nomenclature [2]). Stimuli included natural sounds, ambiences, and sound design. The authors found that as the number of loudspeakers within a given reproduction layout increased, subjective preference ratings increased accordingly.

As part of their research into recording orchestral music for 9+10+3 reproduction, a subset of the current authors examined whether listeners could discriminate between immersive orchestral music reproduction conditions including or excluding bottom channels [17]. Results from that study showed that subjects could discriminate between these two playback conditions with a significant success rate of 69% across three different musical excerpts.

The 3D audio stimuli used for the experiments within the current study (SECS. 2.1 and 2.2) were also used for a concurrent study that attempts to quantify possible objective differences between reproduction conditions with and without the bottom layer of loudspeakers [1]. Playback of the various stimuli was captured using several measurement microphone techniques from previous research in 2D and 3D audio reproduction across three different acoustic environments. The same three reproduction conditions used in the current study (SEC. 2.3) were compared in terms of a set of acoustic features designed to predict reverberance, clarity, envelopment, and apparent source width, ratios of directional sound energy, and various spectral features. Results indicate a consistent trend toward greater low-frequency energy when playback conditions included the bottom layer of loudspeakers [1].

Several reasons for this phenomenon are suggested. First, some instruments, such as a piano or upright bass, tend to radiate low frequencies more efficiently from physically lower areas of the instruments [18], which may be best captured by microphones positioned near the floor. The floor itself may also contribute to a greater radiation of low-frequency information, which may also be best captured by lower-elevated bottom channel-specific microphones, as described in SECS. 2.1 and 2.2. Finally, low-frequency information reproduced from lower-elevated loudspeakers should not suffer from the spectral notches that may be present when sound reproduced from head-height loudspeakers interacts with floor reflections within the reproduction environment [19].

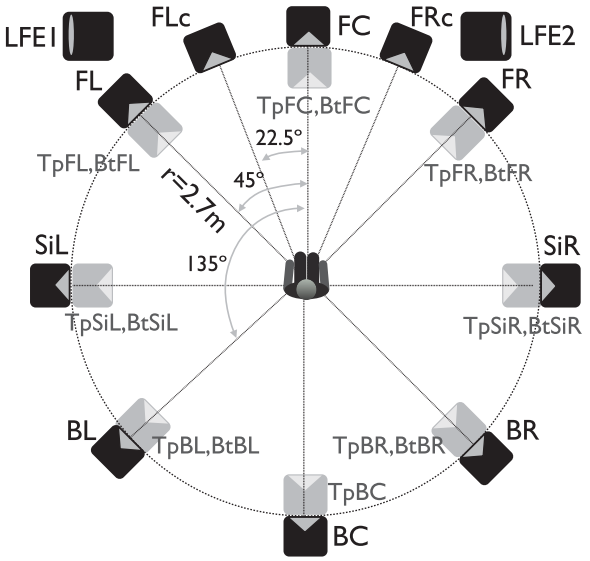


Fig. 1. Tokyo University of the Arts' "Studio B" 9+10+8 loudspeaker layout, as seen from above. Channel naming convention as per [2].

1.2 9+10+8 Audio Reproduction

The 9+10+8 reproduction is identical to 9+10+3 (also known as NHK 22.2, a standardized broadcast format in Japan [5]) in terms of number and spatial positions of loudspeakers but adds bottom channels for the Side Left and Right, and Rear Left, Center, and Right speaker positions (Figs. 1 and 2). For this study, the hope was that an even spatial distribution of loudspeakers across all three vertical layers would aid in providing a more complete understanding of the influence of the bottom-layer within 3D audio reproduction, as opposed to the reproduction formats used in [16] and [17], which only included bottom channels in front of the listener. A complete layer of bottom channels should also help in generalizing the results to current commercial 3D audio formats that include bottom channels both in front of and behind the listener.

2 METHOD

2.1 3D Sound Scenes Under Investigation

Excerpts of 20–30 s from seven 3D sound scenes were selected as material from which the stimuli for this (and the concurrent objective study) were derived. All sound scenes

were recorded specifically or adapted for 9+10+8 reproduction. These recordings were selected in an attempt to cover a range of musical and nonmusical content and different 3D audio production styles within a compact set of stimuli. Extensive documentation and methodological explanations of the production of these recordings can be found in [20, 1] and an online audio/visual repository: https://doi.org/10.5281/zenodo.7563813.

The seven 3D sound scenes are described as follows:

- Rock: up-tempo alternative rock song with a musically and spatially dense arrangement (drums, percussion, bass, acoustic and electric guitars, synthesizers, vocal). 360° perspective combining both "creative" and "re-creative" recording and mixing aesthetics.
- Piano: solo piano performance of a jazz standard recorded in a large studio, using a combination of complex direct and ambient sound microphone arrays; a realistic sonic image of a grand piano.
- Organ: Solo pipe organ recorded at Tokyo University of the Arts' Sōgakudō Concert Hall. A traditional stereo recording setup (Decca tree with outriggers, room mics, and "close" mics) augmented with additional microphones to create a 3D recording.
- Bass: a double bass improvising in a contemporary jazz style. Recorded in a hemi-anechoic environment (Reverberation Time (RT 60) = 0.1 s: a presentation of an acoustic instrument free from interactions between direct and reflected sound.
- Water: an outdoor recording of a waterfall in Takachiho, Japan. Sound captured using a quasi-spherical near-spaced microphone array of 24 DPA 4017c shotgun microphones, designed based on the Boundary Surface Control principal [21].
- Urban Ambience: outdoor ambience recorded at Roy Terrace Community Gardens in Montreal, Canada, using an em32 Eigenmike® from mh acoustics. The excerpt includes the sounds of a passing cyclist, skateboarder, and a child playing. Decoded for 9+10+8 reproduction from a fourth-order Ambisonics b-format audio file, using an open-source AllRAD decoder plugin.
- Taiko: a small taiko drum ensemble (one ōdaiko and two shime-daiko) recorded in a large studio. A simple one microphone per loudspeaker channel setup that yields realistic horizontal and vertical imaging.

2.2 Stimuli Creation

All music recordings were made at 96-kHz/24-bit resolution; outdoor ambience recordings were captured at 48-kHz/24-bit resolution, later sample-rate converted to 96 kHz/24 bit for integration with the other recordings. The music recordings were made by a team of professional music producers/recording engineers with a significant level of previous experience recording and mixing 2D and 3D multichannel audio for commercial release, broadcast, live-sound, and experimental recording sessions. Recordings were mixed or rendered at Tokyo University of the Arts Senju Campus' Studio B (see: SEC. 3.2). The two nonmusical 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, and
- "X": a downmix in which the bottom channel signals have been merged with their corresponding main layer signals (e.g., Bottom Front Centre + Front Centre = Front Centre X) at a 1:1 ratio, with no reproduction from the bottom-layer loudspeakers.

The recordings described in SEC. 2.1 were generally created with an aim to capture sound that had a relevant or "ecologically valid" relationship to the bottom reproduction layer, i.e., sound that normally comes from below the listener in real-life listening is reproduced in the stimuli through loudspeakers situated below the listener. A similar approach was used in the study by Grewe et al. study [16]. Bottom-layer microphones were generally placed within 1 m of the floor, typically with a downward facing angle, and often captured direct and reflected sound and sonic perspectives that would not normally be considered for stereo or 5.1 reproduction or even 3D audio systems without bottom channels.

For example, for the piano recording, three microphones were placed directly underneath the piano, capturing sound that, although likely not useful for main-layer loudspeakers, was reported by the recording team as being valuable in constructing an aesthetically pleasing vertical image of the instrument that seemed to well-represent the complex, directionally dependent timbral profile of the piano [18]. Similarly, for the Taiko recording, the sonic image of the ensemble was created primarily from five close main-layer microphones and three close bottom-layer microphones. The goal was for the combination of these layers to provide the listener with a strong sense of both the horizontal and vertical location of the instruments (the ōdaiko was physically higher than the two shime-daiko) and to give a more complete picture of the complex tonal characteristics of the drums, particularly low-frequencies from the ōdaiko. More details can be found in [20].

When comparing reproduction systems of varying numbers or layers of loudspeakers, one typically begins by recording/mixing a given sound scene for the largest reproduction condition under test and then either actively remixing for all other formats under test, as seen in studies by Francombe et al. [11] and Howie et al. [22], or using a downmixing scheme or algorithm to create additional stimuli, as carried out in studies by Sugimoto et al. [23] and Ando [24].

For the current study, the decision to compare the original 9+10+8 mixes with the fairly simple derivations "Cut" and "X" was based on the aforementioned clear relationship

between bottom-layer microphone signals and reproduction channels. It is always possible that such a method could introduce unwanted spatial or timbral artifacts, particularly for the case of the "X" condition. However, it was felt that this potential was an acceptable tradeoff to avoid the introduction of unwanted variables of technical or aesthetic bias within a human engineer's remixing decisions. Neither solution is ideal, but for the current study, the "passive" method for creation of stimuli seemed the overall better choice.

For the "Rock" example specifically, the method for generating the "Cut" version was found to be somewhat problematic. Within the original 9+10+8 mix, certain microphone signals had been panned to the bottom-layer loudspeakers for primarily spatial-aesthetic reasons, in contrast to the other sound scenes, in which bottom-layer signals typically have a direct physical relationship to the instruments or ambience they are meant to reproduce. This production style was in keeping with the more "creative" or "hyper-realistic" aesthetic found in many commercial pop/rock mixes. When all bottom-layer signals were cut indiscriminately, the balance of the instruments within "Rock's" musical arrangement changed drastically, adding an unwanted variable to the comparison of the three reproduction conditions.

Therefore, a remix was created for the "Cut" version. Though most microphone signals assigned to the bottom layer were still eliminated, a small number of direct-sound microphone signals were remixed to the main-layer loudspeakers to retain continuity within the musical arrangement between all three conditions. It was hoped this would result in a listening comparison more in line with the "Full" vs. "Cut" comparisons for the other sound scenes.

2.2.1 Stimuli Loudness Matching

The playback of each stimulus was recorded using a Brüel & Kjær Type 4128 Head and Torso Simulator situated at the listening position in Studio B. These recordings were made to an Avid Pro Tools audio workstation at 96-kHz/24-bit resolution, with the onboard microphone preamplifiers and analog-digital converters of an RME Fireface UFX+ audio interface. The integrated loudness of each stimulus was measured using a professional software loudness meter plugin (HOFA 4U Meter) set to the EBU +9 scale [25]. Global gain adjustments were then applied to each multichannel audio file until playback of all stimuli was level-matched to within 0.1 LUFS of each other, within each sound scene. This method has been used effectively in several previous studies comparing 3D audio stimuli [22, 26–28]. The levels of each sound scene group were then adjusted relative to each other to make for a comfortable and balanced listening experience throughout the test.

2.3 Listening Tests

Two subjective listening tests were designed to investigate possible perceptual differences between the three reproduction conditions under investigation. The first test measures the ability of subjects to successfully discriminate between the "Full," "Cut," and "X" conditions while eliciting salient perceptual attributes from listeners. The second test asks subjects to compare the three playback conditions for each sound scene based on perceptual attributes collected from the first experiment, as well as general preference.

2.4 Selection of Subjects

For both experiments, subject selection focused on finding "experienced" listeners, whose data would be more consistent or powerful than that of naive listeners [29–34], thereby requiring fewer subjects to achieve meaningful results. Based on previous research into listener performance in 3D audio evaluation [27, 28], a minimum level of 3 years' audio production experience or training was required, unless the subjects had a high level of previous experience with similar listening tests. Musical training, technical ear training, and previous experience hearing 3D audio were also considered to be an asset but not mandatory. The use of experienced listeners in this study is not meant to imply that naive listeners may be unable to perceive differences between the reproduction conditions under test.

2.5 Experimental Venues

Both experiments were performed in three different venues of contrasting physical volumes, acoustical treatments, and design philosophies. For each room, the "listener position" was located at a point equidistant to all main layer loudspeakers (Figs. 1 and 2). The loudspeakers within all three testing venues were visible to subjects, as was common within previous studies related to spatial audio evaluation where acoustically transparent curtains were either not available or were not deemed necessary to obscure the number and position of loudspeakers from the listener [10–12, 22, 23, 16]. Descriptions of each venue follow.

2.5.1 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 conforms to ITU BS.1116 [35] recommendations for critical listening environments except for reverb time, which is somewhat outside of the recommended window (RT 60 = ca. 0.4 s at 500 Hz). Studio B is equipped with 27 KS Digital C5 2-way powered studio monitors. Speaker positions conform to ITU recommendations for 9+10+3 reproduction [2], with the five added bottom channels matching the horizontal angles of their corresponding main-layer speakers. As shown in Figs. 1 and 2, speaker azimuths remain constant across all three layers for vertically associated loudspeaker positions (e.g., TpFL, FL, BtFL), while the angles of elevation for the top and bottom layers mirror each other.

2.5.2 Multichannel Audio Laboratory, Rochester Institute of Technology

The Multichannel Audio Laboratory (MAL) is a multichannel mixing environment located at the Rochester Institute of Technology (floor area = 24.78 m^2, ceiling height

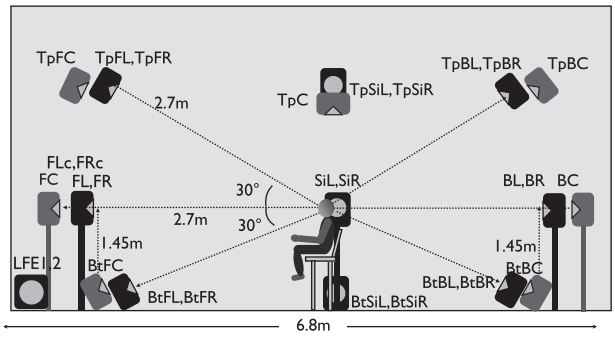


Fig. 2. Studio B 9+10+8 loudspeaker layout, as seen from the side.

= 2.8 m, T30 = 0.37 s). A converted conference room with added acoustical treatment to remove strong wall reflections, this space does not comply with ITU BS.1116 recommendations and is more representative of everyday listening environments. The room is equipped with 27 Genelec 8020B loudspeakers. Speaker positions in MAL are identical to those of Studio B in terms of horizontal azimuths. The angle of elevation for the top layer loudspeakers is +30°, whereas the bottom layer is somewhat shallower than in Studio B, at –20°.

2.5.3 Immersive Media Lab, McGill University

The Immersive Media Lab (IMLab) is a multichannel mixing environment located in the Elizabeth Wirth Music Pavilion, McGill University. The room's walls and ceilings contain a mixture of absorptive, diffusing, and reflective surfaces, with a reflective floor surface (floor area approximately 37 m^2, ceiling height approximately 5 m, RT60 = 0.04 s at 500 Hz). IMLab complies with most ITU recommendations for multichannel playback environments, though some loudspeaker locations are closer to a wall than 1 m, and background noise from in-lab computers somewhat exceeds recommended levels (22 dBa, or approximately NR17). IMLab is equipped with 33 ATC SCM25a Pro powered studio monitors and 3 ATC SCM50ASL Pro powered studio monitors in the FLc, FC, and FRc positions. Unlike Studio B and MAL, the BtFL, BtFR, BtBL, and BtBR loudspeakers are in different horizontal positions from their corresponding main- and height-layer channels. The height-layer loudspeakers have an angle of elevation between 29° and 29.5°, while the bottom-layer loudspeakers have an angle of elevation of –25°.

3 EXPERIMENT 1

3.1 Listening Test

A listening test was conducted to determine how successfully listeners could discriminate between the three reproduction conditions for each sound scene and which perceptual changes between reproduction conditions were most salient to their decision-making process. Subjects were seated at the listening position of their respective testing venue, and it was explained that they would be comparing three different 3D audio reproduction conditions, "Full," "X," and "Cut," across various sound scenes. They were then presented with a simple graphical user interface (GUI) that allowed them to listen to and compare these conditions for each sound scene (all stimuli) for as long as they wished, with the aim to provide all subjects across all three venues with the same baseline understanding of the experimental conditions.

Once this training module was complete, subjects were presented with the GUI for the listening test, and their task was explained verbally, which consisted of comparing stimuli using a simple triad test, implemented with Cycling '74's Max 8 software. For each trial, one of the seven sound scenes was played on a continuous loop. Subjects were instructed to switch between stimuli labeled "A," B," and "C" at their leisure and determine which two were the same (e.g., "1" and "3" are the same condition, while "2" is different). For each trial, playback of all stimuli was time aligned, while micro-crossfades ensured seamless switching between the reproduction conditions. Stimulus assignment to letters "A," "B," and "C" as well as the order of musical excerpts and reproduction condition pairings were randomized within the testing program.

Within each trial, once a decision had been made, the subject was asked to use a text box to provide a term or short phrase related to the aspect of the sound scene that was most salient to their decision (e.g., "vertical image changed," "more bass in B"). There were three possible pairwise comparisons for each sound scene, for a total of 21 trials. Subjects were instructed to set a comfortable listening level during the training phase and to then leave the level unchanged for the remainder of the test. Subjects took an average of 50 min to complete the test, after which they were asked to fill out a short demographic survey. Instructions were provided in English in all testing venues and in English and Japanese for Group 1.

3.2 Results

3.2.1 Participants

A total of 26 subjects participated across the three listening venues, all of whom reported normal hearing and previous experience hearing 3D audio. Table 1 shows the mean and standard deviation for subject demographics per group. (Group 1 = Studio B/Tokyo, Group 2 = MAL/Rochester, Group 3 = IMLab/Montreal). The groups do not show any large or obvious deviations across the included demographic variables.

3.2.2 Discrimination Results

Table 2 shows the discrimination success rates for each comparison of reproduction conditions, across all sound scenes, across each group, with p values using the Bonferroni Correction. As can be seen, it is highly statistically significant that participants across all three groups could discriminate between all of the playback condition comparisons.

Table 1. Subject age and previous experience for Audio Production (AP), 3D Audio production (3D), Musical Training (MT), and Technical Ear Training (TET) per group. Mean and standard deviation are used for numeric variables, and percentage is used for the binary (yes/no) variable.

Group	#	Age	AP	3D	MT	TET
1	9	33 (SD = 14)	11 (SD = 12)	4 (SD = 5)	12 (SD = 3)	78%
2	8	30 (SD = 11)	10 (SD = 9)	5 (SD = 7)	15 (SD = 9)	63%
3	9	32 (SD = 9)	9 (SD = 8)	4 (SD = 4)	13 (SD = 5)	88%

Table 2. Discrimination success rates for each comparison across each group. The p value is calculated using a Binomial Test with the Bonferroni Correction for multiple comparisons and represents where the discrimination rate is statistically significant. Because the listening test was a triad test with three possible responses, the chance of selecting the correct response randomly is 33.3%.

Group	Cut vs. X	X vs. Full	Full vs. Cut
1	73% ($p < 0.001$)	98% ($p < 0.001$)	93% ($p < 0.001$)
2	56% ($p < 0.001$)	82% ($p < 0.001$)	87% ($p < 0.001$)
3	60% ($p < 0.001$)	82% ($p < 0.001$)	76% ($p < 0.001$)

3.2.3 Discrimination Results by Stimuli

Table 3 shows subject discrimination rates for each comparison, for each sound scene, across all three groups. For both the "Full vs. Cut" and "X vs. Full" comparisons, participants appear to be able to discriminate between stimuli across all sound scenes with a high degree of success. The exception is the "Full vs. Cut" comparison for the "Piano" sound scene, which seems to be more difficult for listeners to differentiate. For the "Cut vs. X" comparison, successful discrimination appears to be more tied to specific sound scenes (Bass, Organ, Water), whereas discrimination within "Rock" and "Piano" was somewhat less successful and even less successful for the "Taiko" and "Urban Ambience" sound scenes.

3.2.4 Elicited Salient Auditory Attributes

A total of 612 unique responses were collected across all three listening groups in relation to the most salient difference between the reproduction conditions within each trial. The vast majority of these responses were in English, though some subjects provided responses in Japanese for specific trials. Japanese terms or phrases were translated to English by a researcher familiar with auditory attribute nomenclatures in both languages, confirming these translations with other researchers when necessary. These comments were then searched for terms or synonyms of terms common to subjective spatial audio evaluation. Terms with similar or identical meanings, such as "low end," and "low frequencies," were pooled together, referencing lists of attributes from previous work [36–38] and a thesaurus, a method adopted from [22]. The counts for identical or similar terms (e.g., "envelopment," "enveloping") were then summed, per group.

Table 4. Most common elicited reasons for subject discrimination between reproduction conditions. EIS = Envelopment/Immersion/Spaciousness

Auditory Attribute	Total	Grp 1	Grp 2	Grp 3
Timbre/Tone	131	38	47	46
Low-Frequency Information	116	66	33	17
Vertical Image Change	87	47	10	30
Horizontal Image Change	82	17	23	42
EIS	31	10	6	15
Clarity	23	13	10	0
Localization	15	7	1	7
Phase	13	7	1	5

Table 4 shows the count for the most common auditory attributes collected across all three groups. Attributes related to "timbre and tone," "low frequency information," "vertical image change," and "horizontal image change" remained the most common across all three listener groups, although the rank order of those attributes changed per group. For Group 1, changes in perception of low-frequency content, followed by vertical imaging and timbre seem to be the most useful for subject discrimination between reproduction conditions. For Groups 2 and 3, timbre appears

Table 3. Discrimination success rates for each comparison across each sound scene. UA = Urban Ambience.

Stimuli	Cut vs. X	X vs. Full	Full vs. cut
Bass	96% ($p < 0.001$)	92% ($p < 0.001$)	92% ($p < 0.001$)
Organ	69% ($p < 0.01$)	85% ($p < 0.001$)	88% ($p < 0.001$)
Piano	58% ($p = 0.21$)	86% ($p < 0.001$)	65% ($p = 0.02$)
Rock	60% ($p = 0.12$)	81% ($p < 0.001$)	76% ($p < 0.001$)
Taiko	44% ($p = 1$)	72% ($p = 0.001$)	79% ($p < 0.001$)
UA	46% ($p = 1$)	100% ($p < 0.001$)	100% ($p < 0.001$)
Water	72% ($p < 0.001$)	100% ($p < 0.001$)	100% ($p < 0.001$)

to be the most common factor for discriminating between reproduction conditions, with vertical imaging not as frequently reported as for Group 1.

4 EXPERIMENT 2

4.1 Listening Test

A listening test was implemented to compare the three reproduction conditions of each sound scene in terms of salient perceptual auditory attributes. Based on the listener-elicited responses from Experiment 1 (Table 4) the four following perceptual attributes were chosen:

- **Low Frequency Presence**: amount of presence felt from the low frequencies within the sound scene.
- **Vertical Image Spread**: amount the sound scene or specific sound images spread vertically (i.e., from the floor to the ceiling).
- **Horizontal Image Spread**: amount the sound scene or specific sound images spread horizontally (i.e., width).
- **Naturalness of Timbre**: how natural or realistic is the timbre of the sound scene or specific components of the sound scene.

The naming and definitions of these terms were designed to be harmonious with the original, highly varied subject responses from Experiment 1. The main exception is "Timbre," which is a relatively multimodal term. Rather than investigate a simple bi-polar comparison of one aspect of timbre, such as "Bright" vs. "Dark," it was decided to examine a higher-order aspect of timbre, "Naturalness of Timbre," a term that proved effective in evaluating perceptual differences between various 3D sound capture methods in two previous studies [26, 39]. "Preference" was also included for investigation to enable a more direct comparison of the current study's findings with results from the study by Grewe et al. [16] (see: SEC. 1.1).

The listening test was implemented using Cycling '74's Max 8 software. Subjects were seated at the listening position of their respective testing venue, then given time to familiarize themselves with the testing interface. Test instructions and definitions of the perceptual attributes being investigated were provided both verbally and in written form. It was explained to the subjects that they would be evaluating three different reproduction conditions for each sound scene and what those conditions represented. For each trial, subjects were asked to evaluate stimuli labeled "1," "2," and "3" for a given attribute, using a set of continuous sliders (0–100). Anchor words were provided at the extremes of each slider. Though this test effectively operated as a "rating" test, the goal was not to determine or prove that any one reproduction condition was "better" or "best" but rather to help characterize and quantify perceptual differences between the stimuli.

Subjects were instructed to treat each trial as a new task and were not required to relate their judgments to how they used the scale within previous trials. Playback of stimuli was time-aligned and continuously looped. The test was administered in blocks of seven trials per perceptual attribute to allow subjects to focus on one aspect of the sound scene at a time, a total of 35 trials. For each trial, stimulus assignments to "1," "2," and "3" were randomized, as was the order of attribute trial blocks. Subjects were instructed to set a comfortable listening level before completing the first trial and then leave the level unchanged for the remainder of the test. Subjects took an average of 50 min to complete the test. Upon completion of the test, any subjects who had not already participate in the Experiment 1 were asked to fill out a short demographic survey.

4.2 Results

4.2.1 Participants

A total of 20 subjects across the three venues participated in the listening test, all of whom reported having normal hearing and previous experience hearing 3D audio reproduction. All subjects had previously participated in Experiment 1, except for one new subject in Group 1 and one new subject in Group 3. Table 5 shows the mean and standard deviation for subject demographics per group. As compared with Experiment 1, the subject groups now have somewhat more pronounced differences between them, particularly in terms of mean age, which decreases from Group 1 (Studio B) to Group 2 (MAL) to Group 3 (IMLab), respectively. The listeners in Group 1 are somewhat more experienced than those in Groups 2 and 3 across all demographic factors except for Musical Training, where Groups 1 and 2 are similar.

4.2.2 Attribute Ratings

Table 6 shows the results of a three-way analysis of variance (ANOVA) fit for each attribute individually. The Condition (Full vs. X vs. Cut), Group (1, 2, or 3), and Sound Scene (piano, taiko, organ, water, bass, urban soundscape, rock) variables were used to predict the attribute Rating. The pairwise interaction effect between each predictor variable was also investigated. For the overall model parameters, the Adjusted R^2 is presented alongside the p value for the model. The p value associated with each predictor variable and interaction is also presented.

The three-way ANOVA models fit to each attribute individually indicate that significant interaction effects between Condition and Group occurred for almost every attribute: the Condition was rated differently depending on which group was doing the rating. The presence of this interaction suggests a value in examining the attribute rating data by Group.

Fig. 3 shows a general trend wherein Groups 2 and 3 are not rating reproduction conditions very differently from each other. For these groups, each attribute appears to be roughly equivalent, regardless of whether the bottom layer signals are on, cut, or downmixed to the main layer loudspeakers. The discrimination data from Experiment 1 (see SEC. 3.2) would suggest that these groups are indeed capable of hearing the difference between the stimuli and that these attribute ratings are a matter of choice and not simply

Table 5. Experiment 2 subject age and previous experience: same demographics and analysis as Table 1.

Group		Age	AP	3D	MT	TET
1	8	41 (SD = 16)	16 (SD = 15)	7 (SD = 8)	18 (SD = 15)	88%
2	6	33 (SD = 11)	12 (SD = 10)	6 (SD = 7)	18 (SD = 6)	67%
3	6	29 (SD = 6)	7 (SD = 5)	4 (SD = 3)	12 (SD = 6)	83%

a function of nondiscrimination. In contrast to Groups 2 and 3, Group 1 rated the "Full" reproduction condition much higher than the "Cut" and "X" Conditions for all attributes except for "Horizontal Image Spread."

The ANOVA models summarized in Table 6 also indicate that significant interaction effects between Condition and Sound Scene occurred for every attribute: depending on which Sound Scene was being evaluated, the ratings of the Conditions changed. The presence of this interaction suggests a value in examining subject attribute ratings by Sound Scene, which are shown in Fig. 4—there are several trends worth noting, per attribute.

For Horizontal Image Spread, the "Full" condition for the "Rock" sound scene was rated particularly high. The

Table 6. Results of a 3-way ANOVA, fit for each attribute individually. For overall model parameters, the Adjusted R^2 is presented with the p value for the model. The p value associated with each predictor variable and interaction is also presented. p values lower than 0.05 are in bold. Acronyms: Horizontal Image Spread (HIS), Vertical Image Spread (VIS), Naturalness of Timbre (NT), Low Frequency Presence (LFP), Preference (Pref).

Attribute	HIS	VIS	NT	LFP	Pref
Adjusted R^2	0.11	0.15	0.19	0.20	0.16
$F(p)$	1.99 ($p < 0.001$)	3.09 ($p < 0.001$)	3.66 ($p < 0.001$)	2.64 ($p < 0.001$)	3.09 ($p < 0.001$)
Condition	0.766	**0.002**	0.153	< **0.001**	0.081
Group	< **0.001**	0.228	0.881	0.700	0.236
Sound Scene	**0.010**	0.603	< **0.001**	0.293	**0.002**
Condition * Group	0.075	< **0.001**	< **0.001**	< **0.001**	< **0.001**
Condition * Sound Scene	< **0.001**	**0.004**	< **0.001**	< **0.001**	< **0.001**
Group * Sound Scene	0.775	0.779	< **0.001**	0.603	0.104

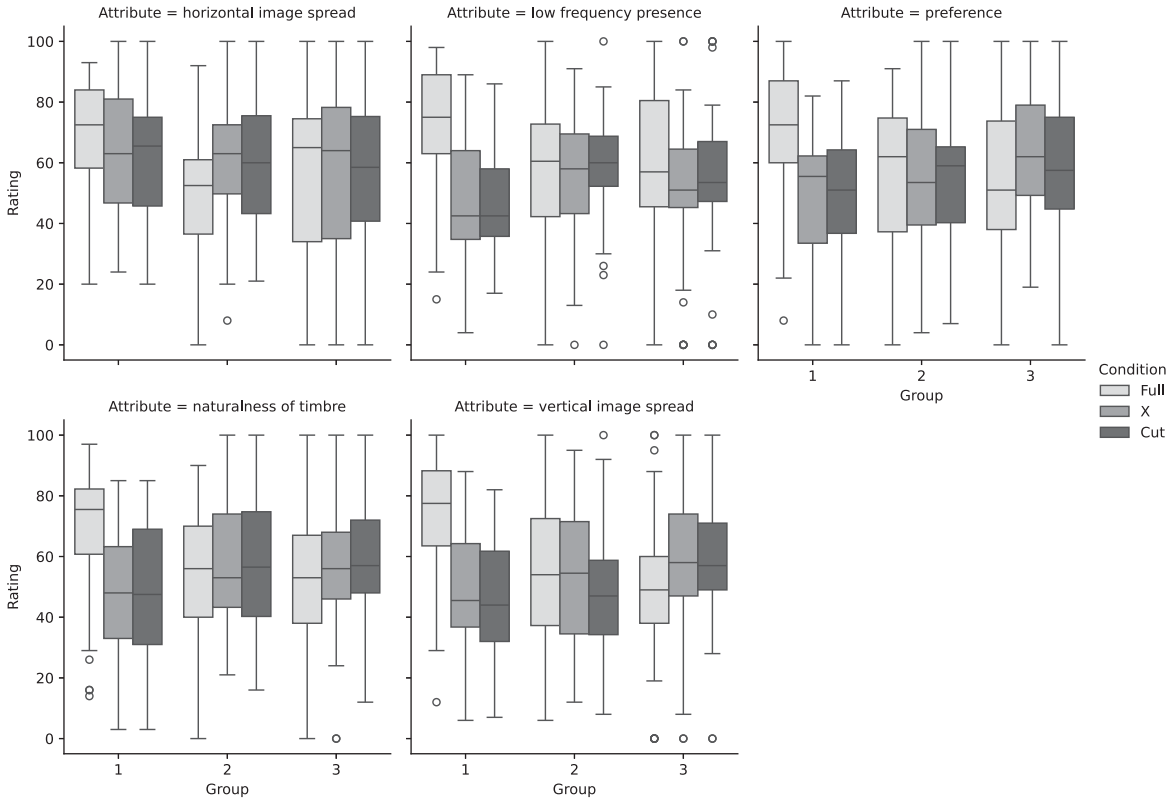


Fig. 3. Subject ratings for each attribute, per group, across all sound scenes.

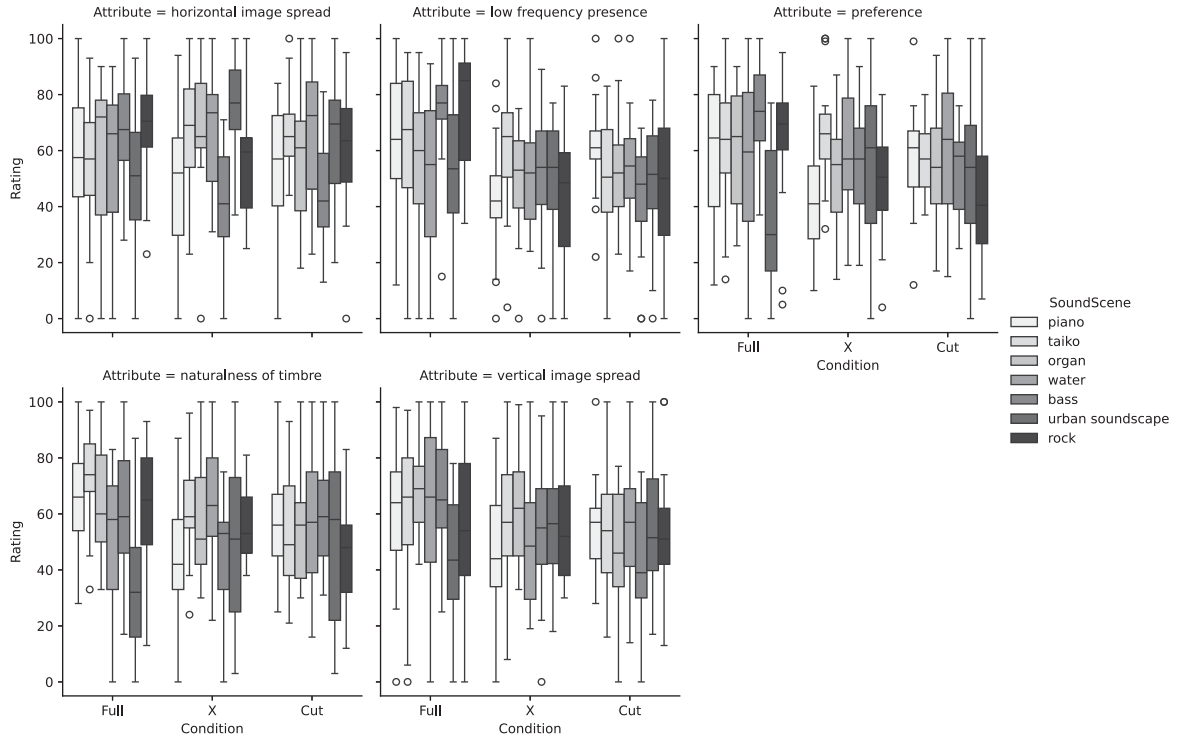


Fig. 4. Subject ratings for each auditory attribute, for each sound scene, across all three listener groups.

"X" condition for "Urban Ambience" was particularly high, whereas the same condition was rated particularly low for "Bass." Similarly, the "Cut" condition was also rated particularly low for the "Bass" sound scene.

For both Vertical Image Spread, and Naturalness of Timbre, the "Full" condition was rated especially low for the "Urban Ambience" sound scene. For the "Piano" sound scene, the "X" condition was rated particularly low for Naturalness of Timbre.

For Low Frequency Presence, the "Full" condition was rated particularly high for "Bass" and "Rock." The "X" condition was rated especially low for the "Piano" sound scene for both Low Frequency Presence and Preference. The "Full" condition was least preferred for the "Urban Ambience" sound scene.

5 DISCUSSION

5.1 Discrimination Between Reproduction Conditions

As shown in Tables 2 and 3, subjects from all three listener groups were able to discriminate between all three reproduction conditions with a high degree of statistical accuracy. This was especially true for the "Full vs. Cut" and "Full vs. X" comparisons, where the average discrimination rates of all groups were all above 75% accuracy. The high accuracy rate for the "Full vs. Cut" comparison is consistent with results from the previous study by Howie et al., in which listeners displayed an average success rate of 69% ($p = < 0.001$) when discriminating between excerpts of orchestral music with and without bottom channels [17]. It is interesting to note that the average success rates in this study are higher, across all three listener groups, which may be due to the 3D audio material used, the use of more bottom-layer loudspeakers (8 vs. 3), listener experience, or a combination of all these factors.

It is somewhat surprising that listeners could discriminate between the "Full" and "X" conditions with such a high degree of accuracy. Recall that these two conditions contain the same signals, the only difference being that for the "X" condition, the bottom-layer loudspeaker signals were combined with the main-layer speaker signals in a 1:1 downmix, with no sound being reproduced from the bottom-layer speakers. This indicates that negative-elevation vertical panning and imaging significantly contributed to perceptual differences between otherwise identical 3D sound scenes.

Interestingly, listeners had more difficulty discriminating between the "Cut" and "X" conditions, even though these two conditions have different combinations of signals, i.e., represent different "mixes." It appears that within this study, differences related to how identical mixes of the sound scenes were being reproduced (more or less vertical speaker layers) were more obvious to listeners than changes within the mixes of the sound scenes. Perhaps listeners are demonstrating a conscious or unconscious understanding that certain sounds "belong" in certain vertical locations.

Physical factors may also be involved: Cabrera and Tilley, for example, present the argument that low frequencies are reproduced more efficiently from floor-level loudspeakers, as sound reproduced from head-level drivers may suffer from low-frequency notches caused by the first-floor reflection [19]. Fig. 5, taken from [1], shows the combined mean power spectra for all sound scenes, as measured in Studio B with a DPA 4006 omnidirectional microphone. It can be seen that the "Full" playback condition con-

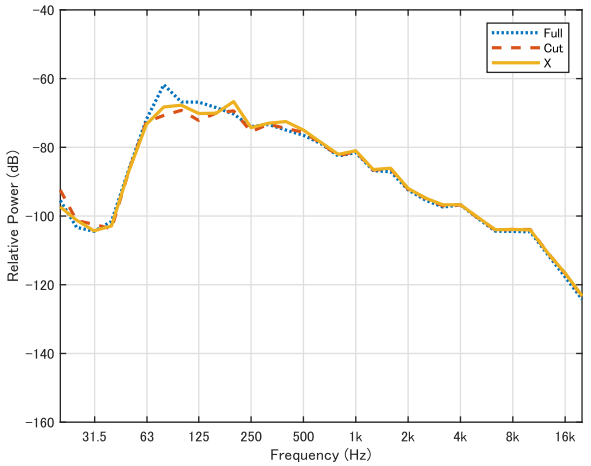


Fig. 5. Combined mean power spectra for all sound scenes, Studio B, reproduced from [1].

tains more measured low-frequency content than the "X" or "Cut" conditions.

As shown in Table 3, sound scene content had some effect on listener discrimination ability. For both the "Full vs. Cut" and "Full vs. X" comparisons, listeners could easily discriminate between conditions across all sound scenes. The exception was "Full vs. Cut" for the "Piano" sound scene, which appears to be a more difficult comparison to make; for this particular sound scene, the addition of lower-elevated sound may have been more subtle than in the other sound scenes, leading to less pronounced perceptual differences such as timbre or vertical imaging.

This result is somewhat in opposition to anecdotal observations from the production team responsible for the piano recording, as well as various individuals who have heard the piano recording as a part of informal listening sessions. Among those listeners, it was generally felt that the bottom channel microphones, especially those near the piano, made an important contribution to the overall presence and realism of the recording. Perhaps the design of the current study was not effective at emphasizing or revealing those differences, or perhaps these listeners were suffering from a certain amount of expectation bias. For the "Cut vs. X" comparisons, discrimination success becomes much more dependent on the sound scene: the average success rates across all three listener groups are only statistically significant within the "Bass," "Organ," and "Water" examples.

5.2 Perceptual Differences Between Reproduction Conditions

5.2.1 Experiment 1

As shown in Table 4, the most common perceptual differences between reproduction conditions observed by listeners within Experiment 1 were related to timbre/tone, low-frequency information, and changes in the vertical and horizontal images and that these were the four most common reasons or attributes for discrimination across all three groups, although changes within vertical imaging seem to be more important or apparent to Group 1 listeners than for those in the other two groups. Differences related to envelopment/immersion/spaciousness were also common across all three groups, although reported comparatively less frequently. These results are somewhat comparable with those of a previous study by a subset of the current authors [22] examining listener discrimination between 9+10+3, 4+7+0, 3+7+0, and 4+5+0 audio reproduction formats. In that study, differences related to vertical imaging and apparent source width, a similar concept to "horizontal image spread," were mentioned by many listeners as being important factors in discriminating between the various formats, with the vertical imaging being a primary cue for when the bottom channels (9+10+3) were active.

5.2.2 Experiment 2

Experiment 2 shows a more distinct difference between responses from Group 1 as compared with Groups 2 and 3. As shown in Fig. 3, for Group 1, the "Full" reproduction condition occupies a different perceptual space than the other two conditions for all auditory attributes under investigation except "horizontal image spread." For those listeners, the inclusion of sound reproduction from the bottom layer resulted in sound scenes that had a greater degree of vertical spread, a more natural timbre, and a greater presence of low-frequency content.

In contrast, for Groups 2 and 3, the three reproduction conditions appear to occupy a similar perceptual space for all attributes, including preference. It is clear from the results of the first experiment that listeners in Groups 2 and 3 are capable of hearing differences between the reproduction conditions, yet in Experiment 2, those same differences do not appear to be strong enough to compellingly influence ratings across the various attributes under investigation. Given that those attributes were derived from almost identical listener groups in Experiment 1, the reason for this difference in perception between Group 1 and Groups 2 and 3 may be due to demographic or room-related factors to be discussed in SEC. 5.4.

Fig. 4 shows that sound scene appears to have an effect on perception between reproduction conditions. There are several examples worth noting. For the "Piano" sound scene, the downmixed version resulted in an unnatural timbre with less low-frequency presence, which listeners seemed to dislike. For both the "Rock" and "Bass" sound scenes, the inclusion of the bottom-layer loudspeakers results in a wider sound image and an increase in low-frequency presence. This is interesting, as there is not an obvious relationship between lower-elevated sound reproduction and an increase in perceived width of sound images. A similar but less-pronounced trend in more low-frequency presence is also observed for the "Taiko" and "Organ" sound scenes when the bottom-layer is active. The "Bass," "Organ," "Taiko," and "Rock" sound scenes all contain musical instruments that produce a large amount of low-frequency content or whose musical ranges tend to sit within the area of lower frequencies. This may be why differences in low-frequency perception were more pronounced for these sound scenes than for the others.

5.3 Preference Between Reproduction Conditions

Differences in preference ratings between reproduction conditions were only observed within Group 1's data. Those listeners preferred the "Full" condition across all sound scenes, except "Urban Ambience," where the "Full" condition is preferred less than the "Cut" and "X" conditions. This indicates a general preference for 3D audio reproduction including the bottom-layer within this listener group, as was also the case in the study by Grewe et al. [16]. It also shows that this trend in preference holds true for musical sound scenes, which were absent within [16].

5.4 Group Differences

The three listener groups for Experiment 1 are quite similar across all measured demographics (Table 1), yet these groups display some differences in terms of successful discrimination rates (Table 2) and what aspects of the auditory environment they reported as being most salient between reproduction conditions (Table 4). For Experiment 2, where only Group 1 listeners reported strong perceptual differences between the stimuli for the various attributes under investigation, including general preference between reproduction conditions, there is an increase in variation of experience between the three listener groups, yet all show high levels of audio production experience, 3D audio experience, and musical training (Table 5).

One possible demographic explanation for variation in perception and preference between these groups is familiarity with 3D audio content including bottom channels, which was not specifically measured within the post-test survey. A number of the subjects at Tokyo University of the Arts (Group 1) already had previous experience either creating or listening to content including including bottom-layer reproduction (9+10+3 has been a standardized broadcast format in Japan for over a decade). Conversely, the concept of significant bottom channel content would have been a relatively new concept for many of the listeners in Groups 2 and 3.

Another explanation for the differences in results between listener groups, particularly for Experiment 2, comes from an examination of room-related factors. First, although the angle of elevation of the height channels is essentially the same for all three testing venues (29° to 30°), the negative angle of elevation for the bottom layer loudspeakers for both the Group 2 (–20°) and Group 3 (–25°) venues is shallower than for the Group 1 venue (–30°). This difference in vertical speaker displacement may help to explain why differences in vertical imaging were more relevant or perceivable to Group 1.

For Group 3, the FLc, FRc, and FC loudspeakers were significantly larger than the other main-layer loudspeakers in the venue, physically extending closer to the bottom layer loudspeakers, further decreasing the angular difference between the vertical layers for those three channels. The musical sound scenes in the current study all feature important elements in front of the listener, with most mixes making extensive use of the FLc, FC, and FRc channels. For these types of sound scenes, a significant asymmetry between vertical speaker layers may reduce perceptual differences between reproduction conditions, particularly for changes in vertical imaging.

At the Group 1 testing venue, a 1-m-wide ring of absorptive material was strategically placed to somewhat dampen the first-floor reflection from the main-layer loudspeakers. Although the investigators observed that this created a relatively small perceptual change, it may have provided listeners a greater clarity in understanding the differences between reproduction conditions. It was not possible to install similar acoustical treatment within the Group 2 or 3 venues. A future experiment designed to investigate how listening venue factors such as vertical speaker layer angle or room acoustic conditions affect listener perception of bottom-layer sound reproduction would be valuable.

There is also some anecdotal evidence to support the influence of the room-reproduction chain on the results of this study. Three of the subjects from Group 2 had the opportunity to visit the Group 1 testing venue and compare the three reproduction conditions across the various sound scenes: all three agreed that perceptual differences between conditions, particularly vertical imaging, were more obvious in the Group 1 venue than in the Group 2 venue.

6 CONCLUSION

This study investigated perceptual differences and listener preference between 3D audio reproduction with and without bottom channels. Three different reproduction conditions were compared: "Full," the original mix, using all three vertical speaker layers; "X," the bottom-channel signals downmixed to the main layer loudspeakers; and "Cut," all bottom-channel content muted. Stimuli were derived from 20–30-s excerpts of a range of musical and nonmusical sound scenes created for a 9+10+8 3D audio reproduction environment using a range of 3D audio recording and mixing techniques. Two different experimental listening tests were executed in three different listening venues with three different listener groups. The first test investigated subjects' ability to discriminate between the three reproduction conditions and to ascertain what aspects of the sound scene listeners found most relevant when comparing these reproduction conditions. The second test asked subjects to rank the reproduction conditions based on four key auditory attributes derived from the first test, as well as general preference. Analysis of listener data revealed the following:

1) Listeners across all three testing venues could discriminate between the three reproduction conditions with a very high degree of accuracy.
2) Listeners had greater success discriminating between the "Full vs. Cut" and "Full vs. X" conditions, than the "Cut vs. X" conditions.
3) Results for the "Full vs. Cut" comparison are consistent with a previous study comparing 3D orchestral music reproduction with and without bottom channels [17].

4) Across all three listener groups, the most salient aspects of the sound scene for differentiating between reproduction conditions were changes related to low-frequency energy, horizontal and vertical imaging, and timbre/tone.
5) Listeners in Group 1 (Tokyo University of the Arts) focused more on vertical imaging and low-frequency presence, whereas listeners in Groups 2 (Rochester Institute of Technology) and 3 (McGill University) focused more on changes in timbre and tone.
6) Only listeners in Group 1 showed compelling differences in their ratings of the reproduction conditions in terms of the perceptual attributes under investigation. They found the "Full" condition contributed to a greater sense of vertical image spread, low-frequency presence, and naturalness of timbre as compared with the other reproduction conditions. They also preferred the "Full" condition, which is consistent with results from a previous study by Grewe et al. [16].
7) The differences in perception of the reproduction conditions between the groups are likely related to several factors, including previous experience hearing or creating content for 3D audio systems with bottom channels, the angle of elevation of the bottom layer of loudspeakers in the testing venues, and acoustic conditions within the testing venues.
8) A future study designed to investigate how specific types of previous 3D audio experience, as well as room/reproduction conditions, affect listener perception of 3D audio reproduction would be valuable, as would the inclusion of more material derived from Ambisonics-based recordings.

7 ACKNOWLEDGMENT

This work was supported by the Japan Society for the Promotion of Science, Tokyo University of the Arts, McGill University, Rochester Institute of Technology, and JSPS KAKENHI grant numbers 21H03761, 21H03764, and 21F21745. Thanks to Akira Omoto and Florian Grond for use of their recordings.

8 REFERENCES

[1] W. Howie, A. Marui, T. Kamekawa, and F. Grond, "Objective Comparisons of 3D Audio Reproduction With and Without Bottom Channels," in *Proceedings of the 2023 AES International Conference on Spatial and Immersive Audio* (2023 Aug.), paper 9.

[2] ITU-R, "Advanced Sound System for Programme Production," *Recommendation ITU-R BS.2051-3* (2022 May).

[3] ITU-R, "Performance Requirements for an Advanced Multichannel Stereophonic Sound System for Use With or Without Accompanying Picture," *Recommendation ITU-R BS.1909-0* (2012 Jan.).

[4] A. Roginska and P. Geluso, eds., *Immersive Sound: The Art and Science of Binaural and Multi-Channel Audio* (Routledge, New York, USA, 2018), 1st ed.

[5] K. Hamasaki and K. Hiyama, "Development of a 22.2 Multichannel Sound System," *Broadcast Technol.*, vol. 25, pp. 9–13 (2006 Winter).

[6] 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 Jun.), paper 581.

[7] S. Kaneko, T. Suenaga, H. Akiyama, 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 Jun.), paper 10012.

[8] A. Omoto, S. Ise, Y. Ikeda, K. Ueno, S. Enomoto, and M. Kobayashi, "Sound Field Reproduction and Sharing System Based on the Boundary Surface Control Principle," *Acoust. Sci. Technol.*, vol. 36, no. 1, pp. 1–11 (2015 Jan.). https://doi.org/10.1250/ast.36.1.

[9] Y. Tanabe, G. Yamauchi, A. Marui, and T. Kamekawa, "Tesseral Array for Group Based Spatial Audio Capture and Synthesis," in *Proceedings of the 2020 AES International Conference on Audio for Virtual and Augmented Reality* (2020 Aug.), paper 2-7.

[10] K. Hamasaki, T. Nishiguchi, K. Hiyama, and R. Okumura, "Effectiveness of Height Information for Reproducing Presence and Reality in Multichannel Audio System," presented at the *120th Convention of the Audio Engineering Society* (2006 May), paper 6679.

[11] 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 Mar.). https://doi.org/10.17743/jaes.2016.0071.

[12] 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 Jul.).

[13] 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 May), paper 8334.

[14] 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 Oct.), paper 9003.

[15] J. Paterson and H. Lee, eds., *3D Audio* (Routledge, New York, USA, 2022), 1st ed.

[16] Y. Grewe, A. Walther, and J. Klapp, "Evaluation on the Perceptual Influence of Floor Level Loudspeakers for Immersive Audio Reproduction," presented at the *147th Convention of the Audio Engineering Society* (2019 Oct.), paper 10276.

[17] W. Howie, R. King, and D. Martin, "A Three-Dimensional Orchestral Music Recording Technique, Optimized for 22.2 Multichannel Sound," presented at the *141st Convention of the Audio Engineering Society* (2016 Sep.), paper 9612.

[18] J. Meyer, *Acoustics and the Performance of Music* (Springer, New York, USA, 2009), 5th ed.

[19] D. Cabrera and S. Tilley, "Vertical Localization and Image Size Effects in Loudspeaker Reproduction," in *Proceedings of the AES 24th International Conference: Multichannel Audio, The New Reality* (2003 Jun.), paper 46.

[20] W. Howie, T. Kamekawa, and M. Morinaga, "Case Studies in Music Production for 3D Audio Reproduction With Bottom Channels," in *Proceedingsof the AES 2023 International Conference on Spatial and Immersive Audio* (2023 Aug.), paper 3.

[21] A. Omoto and H. Kashiwazaki, "Hypotheses for Constructing a Precise, Straightforward, Robust and Versatile Sound Field Reproduction System," *Acoust. Sci. Technol.*, vol. 41, no. 1, pp. 151–159 (2020 Jan.). https://doi.org/10.1250/ast.41.151.

[22] W. Howie, R. King, and D. Martin, "Listener Discrimination Between Common Channel-Based 3D Audio Reproduction Formats," *J. Audio Eng. Soc.*, vol. 65, no. 10, pp. 796–805 (2017 Oct.). https://doi.org/10.17743/jaes.2017.0030.

[23] T. Sugimoto, S. Oode, and Y. Nakayama, "Downmixing Method for 22.2 Multichannel Sound Signal in 8K Super H-Vision Broadcasting," *J. Audio Eng. Soc.*, vol. 63, nos. 7/8, pp. 590–599 (2015 Jul.) https://doi.org/10.17743/jaes.2015.0062.

[24] A. Ando, "Conversion of Multichannel Sound Signal Maintaining Physical Properties of Sound in Reproduced Sound Field," *IEEE Trans. Audio Speech Lang. Process.*, vol. 19, no. 6, pp. 1467–1475 (2011 Aug.). https://doi.org/10.1109/TASL.2010.2092429.

[25] EBU, "Loudness Normalization and Permitted Maximum Level of Audio Signals," *Recommendation EBU R 128-2020* (2020 Aug.).

[26] W. Howie, D. Martin, D. H. Benson, J. Kelly, and R. King, "Subjective and Objective Evaluation of 9ch Three-Dimensional Acoustic Music Recording Techniques," in *Proceedings of the 2018 AES International Conference on Spatial Reproduction - Aesthetics and Science* (2018 Jul.), paper P10-1.

[27] W. Howie, D. Martin, S. Kim, T. Kamekawa, and R. King, "Effect of Audio Production Experience, Musical Training, and Age on Listener Performance in 3D Audio Evaluation," *J. Audio Eng. Soc.*, vol. 67, no. 10, pp. 782–894 (2019 Oct.). https://doi.org/10.17743/jaes.2019.0031.

[28] W. Howie, D. Martin, S. Kim, T. Kamekawa, and R. King, "Effect of Skill Level on Listener Performance in 3D Audio Evaluation," *J. Audio Eng. Soc.*, vol. 68, no. 9, pp. 628–637 (2020 Sep.). https://doi.org/10.17743/jaes.2020.0050.

[29] F. E. Toole, "Subjective Measurements of Loudspeaker Sound Quality and Listener Performance," *J. Audio Eng. Soc.*, vol. 33, nos. 1/2, pp. 2–32 (1985 Feb.).

[30] S. Bech, "Selection and Training of Subjects for Listening Tests on Sound-Reproducing Equipment," *J. Audio Eng. Soc.*, vol. 40, nos. 7/8, pp. 590–610 (1992 Jul.).

[31] S. Olive, "Differences in Performance and Preference of Trained Versus Untrained Listeners in Loudspeaker Tests: A Case Study," *J. Audio Eng. Soc.*, vol. 51, no. 9, pp. 806–825 (2003 Sep.).

[32] A. Gabrielsson and H. Sjögren, "Perceived Sound Quality of Sound-Reproducing Systems," *J. Acoust. Soc. Amer.*, vol. 65, no. 4, pp. 1019–1033 (1979 Apr.). https://doi.org/10.1121/1.382579.

[33] M. C. Killian and T. M. Tillman, "Evaluation of High-Fidelity Hearing Aids," *J. Speech Hear. Res.*, vol. 25, no. 1, pp. 15–25 (1982 Mar.). https://doi.org/10.1044/jshr.2501.15.

[34] N. Schinkel-Bielefeld, N. Lotze, and F. Nagel, "Audio Quality Evaluation by Experienced and Inexperienced Listeners," *Proc. Meet. Acoust.*, vol. 19, no. 1, paper 060016 (2013 Jun.). https://doi.org/10.1121/1.4799190.

[35] ITU-R, "Methods for the Subjective Assessment of Small Impairments in Audio Systems," *Recommendation ITU-R BS.1116-3* (2015 Feb.).

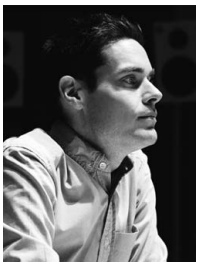
[36] F. Rumsey, "Spatial Quality Evaluation for Reproduced Sound: Terminology, Meaning, and a Scene-Based Paradigm," *J. Audio Eng. Soc.*, vol. 50, no. 9, pp. 651–666 (2002 Sep.).

[37] S. Bech and N. Zacharov, *Perceptual Audio Evaluation – Theory, Method and Application* (John Wiley & Sons, Chichester, UK, 2006).

[38] S. Choisel and F. Wickelmaier, "Evaluation of Multichannel Reproduced Sound: Scaling Auditory Attributes Underlying Listener Preference," *J. Acoust. Soc. Am.*, vol. 121, no. 1, pp. 338–400 (2007 Jan.). https://doi.org/10.1121/1.2385043.

[39] W. Howie, D. Martin, T. Kamekawa, J. Kelly, and R. King, "Comparing Immersive Sound Capture Techniques Optimized for Acoustic Music Recording Through Binaural Reproduction," presented at the *150th Convention of the Audio Engineering Society* (2021 May), paper 10455.

THE AUTHORS


Will Howie


Denis Martin


Atsushi Marui


Toru Kamekawa


Sungyoung Kim


Aybar Aydin


Richard King

Will Howie is an award-winning Recording Engineer and Music Producer, and Researcher and Educator specializing in multichannel immersive audio. He holds a Ph.D. in Sound Recording from McGill University and was a Japan Society for the Promotion of Science postdoctoral research fellow at Tokyo University of the Arts, where his work focused on exploring human perception of 3D audio reproduction. Will has developed and written about numerous 3D audio music production techniques.

•

Denis Martin is an Assistant Professor of Music Technology and Digital Media at the University of Toronto. Previously, he taught at York University, McGill University, and The State University of New York in Potsdam. His research is focused on open audio engineering resources and the perception, operation, and design of dynamic range compressors. In addition, he serves as a collaborator in audio production research that requires expertise in statistical data analysis and experiment design.

•

Atsushi Marui received an M.Sci. degree from Pennsylvania State University and pursued his doctoral studies at The University of Aizu, Japan, from which he received a Ph.D. degree in Computer Science and Engineering. He was also a doctoral student in Sound Recording at the Schulich School of Music, McGill University, where he finished all but dissertation. He is currently a professor at the Faculty of Music, Tokyo University of the Arts, Japan. His research interests include signal processing algorithms and their psychological evaluations of digital audio effect processors, especially on nonlinear distortion processors for musical instruments, artificial reverberation, human–computer interaction, and auditory display.

•

Toru Kamekawa is a researcher, educator, and recording/mixing engineer with a wide range of knowledge in various scientific and artistic disciplines. He studied acoustics at Kyushu Institute of Design and joined the Japan Broadcasting Corporation (NHK) as a sound engineer in 1983. In 2002, Toru joined the Tokyo University of the Arts as an associate professor in the Department of Musical Creativity and the Environment and has been a full professor since 2010. His current research focuses on the relationship between "spatial impressions" and 3D audio recordings and rendering environments, such as 22.2 multi-channel audio and higher order ambisonics.

•

Sungyoung Kim received a B.S. degree from Sogang University, Korea, and Master of Music and Ph.D. from McGill University, Canada. Currently, he works for the Korea Advanced Institute of Science and Technology (KAIST) and Rochester Institute of Technology (RIT) as an associate professor. His research interests are rendering and perceptual evaluation of spatial audio, digital preservation of aural heritage, and auditory training for hearing rehabilitation.

•

Aybar Aydin is a Ph.D. candidate at the McGill University Sound Recording Area. His currents research interests are spatial audio, digital signal processing, and multichannel audio effects. Aybar is a student member at the Audio Engineering Society.

•

Richard King is an Educator, Researcher, and a Grammy Award–winning recording engineer (including Best Engineered Album in both the Classical and Non-Classical categories). Richard is an Associate Professor at McGill University in Montréal, Canada, and as a long-standing member and Fellow of the Audio Engineering Society, he is a regular convention presenter and workshop panelist. His research interests and publications include small environment acoustics, the process of music mixing, and immersive audio recording and reproduction.