H. Lee and D. Johnson, "3D Microphone Array Comparison: Objective Measurements"
J. Audio Eng. Soc., vol. 69, no. 11, pp. 871–887, (2021 November).
DOI: https://doi.org/10.17743/jaes.2021.0038

3D Microphone Array Comparison: Objective Measurements

HYUNKOOK LEE,* *AES Fellow,* **AND DALE JOHNSON**
(h.lee@hud.ac.uk) (d.s.johnson2@hud.ac.uk)

Applied Psychoacoustics Laboratory (APL), University of Huddersfield, Huddersfield, United Kingdom

This paper describes a set of objective measurements carried out to compare various types of 3D microphone arrays, comprising OCT-3D, PCMA-3D, 2L-Cube, Decca Cuboid, Eigenmike EM32 (i.e., spherical microphone system), and Hamasaki Square with 0-m and 1-m vertical spacings of the height layer. Objective parameters that were measured comprised interchannel and spectral differences caused by interchannel crosstalk (ICXT), fluctuations of interaural level and time differences (ILD and ITD), interchannel correlation coefficient (ICC), interaural cross-correlation coefficient (IACC), and direct-to-reverberant energy ratio (DRR). These were chosen as potential predictors for perceived differences among the arrays. The measurements of the properties of ICXT and the time-varying ILD and ITD suggest that the arrays would produce substantial perceived differences in tonal quality as well as locatedness. The analyses of ICCs and IACCs indicate that perceived differences among the arrays in spatial impression would be larger horizontally rather than vertically. It is also predicted that the addition of the height channel signals to the base channel ones in reproduction would produce little effect on both source-image spread and listener envelopment, regardless of the array type. Finally, differences between the ear-input signals in DRR were substantially smaller than those observed among microphone signals.

0 INTRODUCTION

0.1 Background

Three-dimensional (3D) audio is rapidly becoming a new standard for audio content production, delivery, and reproduction. New 3D reproduction formats (e.g., [1–5] and audio codecs (e.g., [6]) are being adopted widely in consumer products as well as streaming and broadcasting services. This is also boosting developments of new techniques and tools for 3D audio content creation. In the context of acoustic recording, a number of 3D microphone array techniques have been proposed over the recent years [7–17]; a comprehensive review of existing 3D microphone arrays is provided in [18]. Furthermore, with the burgeoning interest in head-tracked binaural audio for extended reality applications, Ambisonic microphone systems (e.g., [19–22]) are used more widely than in the past for its convenience in sound field rotation.

As an increasing number of 3D acoustic music recordings are being made, there arises the need for evaluating the qualities of 3D microphone systems used to record them in a systematic way. Much research has been undertaken on the quality evaluation of horizontal-only surround sound recording (e.g., [23–25]). However, the relationship between the 3D microphone setup including the so-called 'height' channels and the perceived qualities of the resulting recordings have not been fully investigated yet.

The present paper is concerned with the evaluations of 3D microphone "array" techniques. In contrast with techniques that place multiple spot microphones close to sound sources, array techniques attempt to capture an acoustic scene using microphones that are arranged in a certain configuration, which is either physically or perceptually motivated [18]. The choice of microphone technique would depend on the recording engineer's philosophy. While some engineers might solely rely on the spatial mixing of spot microphones, some others might use a main microphone array only, aiming to place it at an optimal position to capture both direct sounds and ambience with a desired balance. In many cases, however, one might use both main array and spot microphones for flexibility in post-production. Whichever approach is preferred, it would be essential to capture the reflections and reverberation of a recording space to feed the

*To whom correspondence should be addressed e-mail: h.lee@hud.ac.uk

height and surround channels in order to deliver a realistic and enveloping listening experience in 3D reproduction.

0.2 3D Microphone Array Comparison (3D-MARCo) Database

Several studies have compared different 3D microphone array techniques (see Sec. 2 in [18] for a review). They generally suggest that different techniques had different pros and cons depending on the tested attributes. However, as pointed out in [18], the previous studies had limitations in terms of the number of techniques compared, consistency in the microphone models used for different arrays, and data analysis method. To allow for a more systematic and comprehensive investigation into the perceptual characteristics of different 3D microphone arrays, it would first be necessary to create various types of sound sources recorded using a number of different arrays simultaneously. Furthermore, the microphones and preamps to be used should ideally be of the same manufacturer and brand in order to minimize the influence of recording systems, which would allow for a more controlled comparison on microphone-array-dependent spatial and timbral qualities.

Such a database, named '3D-MARCo' (3D Microphone Array Recording Comparison), has recently been created by the present authors [26, 27]. The recordings were made in a reverberant concert hall using a total of 65 individual microphones, 51 of which were of an identical manufacturer and brand (DPA d:dicate series), as well as first-order and higher-order Ambisonic microphones systems. Using the individual microphones, six different nine-channel or eight-channel spaced microphone arrays were configured. Additional microphones for side, side height, overhead, and floor channels were also used for a possible extension to a larger reproduction format. Five different types of musical performances, comprising string quartet, piano trio, organ, and a cappella singers, were recorded using all of the microphones simultaneously. Furthermore, multichannel room impulse responses were captured for 13 different source positions using all of the microphones to allow for the objective analyses of the microphone arrays as well as the creation of virtual sound sources for future experiments.

0.3 Research Aim

As the first step toward a series of planned formal evaluations of the 3D microphone arrays included in the 3D-MARCo database, the present study aims to provide objective insights into differences among the microphone arrays through the computations of various objective parameters. The results of this investigation will not only serve as bases for explaining perceptual differences among the arrays, which will be determined in future subjective listening tests, but also provide useful practical implications on the choice and use of microphone arrays for different purposes.

The rest of the paper is organized as follows. Sec. 1 briefly summarizes microphone arrays compared in this study. Sec. 2 describes the objective parameters and the methods used for computing them. Sec. 3 then presents and discusses the results.

Table 1. Microphone/loudspeaker channels and labels and the positions of loudspeakers used in the present study.

Channels	Labels	Azimuth (°)	Elevation (°)
Front Left	FL	+30	0
Front Right	FR	–30	0
Front Center	FC	0	0
Rear Left	RL	120	0
Rear Right	RR	–120	0
Front Left height	FLh	+45	+45
Front Right height	FLh	–45	+45
Rear Left height	RLh	+135	+45
Rear Right height	RRh	–135	+45

1 MICROPHONE ARRAYS AND RECORDING SETUP

A total of seven different microphone arrays from the 3D-MARCo database were compared in the present study. This section briefly describes the array configurations. Full details about the database are available in [26, 27]. This paper uses the channel labels and loudspeaker positions presented in Table 1. The azimuth and elevation angles of the loudspeakers were chosen based on ITU-R BS.2051-2 [28]. This configuration is also in line with typical loudspeaker layouts for nine-channel 3D home-cinema systems, such as Dolby Atmos 5.1.4 and Auro-3D 9.1.

1.1 Microphone Arrays

Table 2 lists and categorizes the microphone arrays from 3D-MARCo that were compared in this study. They were chosen for their distinct differences in terms of design concept, physical configuration, and purpose. The physical configurations of the arrays are illustrated in Fig. 1. Detailed information on the microphone models, polar patterns, and microphone angles chosen for each array can be found in [26].

1.1.1 OCT-3D

OCT-3D (Fig. 1(a)), proposed by Theile and Wittek [7], augments the Optimized Cardioid Triangle (OCT)-surround five-channel microphone array [29] with four upward-facing supercardioid microphones placed 1 m above the base layer. The main design goal of OCT is to minimize interchannel crosstalk (ICXT) for accurate frontal image localization. The front triplet uses a cardioid center microphone placed 8 cm in front of the array's base point and two sideward-facing supercardioid microphones, the spacing of which can be varied depending on the desired stereophonic recording angle (SRA). In the 3D-MARCo recording session, a 70-cm spacing was used to produce the SRA of 115° [30]. The rear microphones were backward-facing cardioid microphones with 1-m spacing, placed at 40 cm behind the front supercardioid microphones. In the original OCT-3D proposal [7], the height layer microphones are placed directly above the base layer microphones apart

Table 2. 3D microphone arrays compared in the current study, classified according to [18].

	Perceptually Motivated		Physically Motivated
Main Array	Horizontally and Verically Spaced (HVS) OCT-3D 2L-Cube Decca Cuboid	Horizontally spaced / Vertically coincident (HSVC) PCMA-3D	Horizontally and Vertically Coincident (HVC) Eigenmike EM-32
Ambience Array	Hamasaki Square (HS) with height layer at 1 m above	Hamasaki Square (HS) with height layer at 0 m	

from the front center one. However, in the 3D-MARCo session, the height layer was modified to be of a 1 m x 1 m square to be consistent with the PCMA-3D's height layer.

1.1.2 PCMA-3D

The PCMA-3D (Fig. 1(b)) is based on the 'Perspective Control Microphone Array' design concept [8], which allows a flexible rendering of per-ceived distance in five-channel surround recording. PCMA employs a coincident pair of microphones at each point in the array. By changing the mixing ratio of forward and backward-facing cardioid microphones, a source-to-ambience ratio can be controlled, thus changing the perceived distance of the sound image. This concept has been adapted for PCMA-3D based on three main research findings: (i) vertical microphone spacing (i.e., vertical interchannel decorrelation) did not have a significant effect on perceived spatial impression in 3D sound reproduction [8], (ii) vertical interchannel time difference is an unstable cue for vertical phantom imaging [31], (iii) in order to avoid an unwanted upward-shifting of a source image, the level of the direct sound in each height microphone (i.e., ICXT) should be at least 7 dB lower than that in the corresponding microphone of the base layer [32]. This becomes the basis of the horizontally spaced and vertically coincident (HSVC) array design concept. The 3D-MARCo session used supercardioid capsules for the height layer of PCMA-3D, and they were angled directly upward in order to suppress the ICXT maximally.

1.1.3 2L-Cube

2L-Cube is a technique developed by Lindberg [9]. It employs nine omni-directional microphones in a 1 m x 1 m x 1 m cube arrangement, thus mainly relying on interchan-

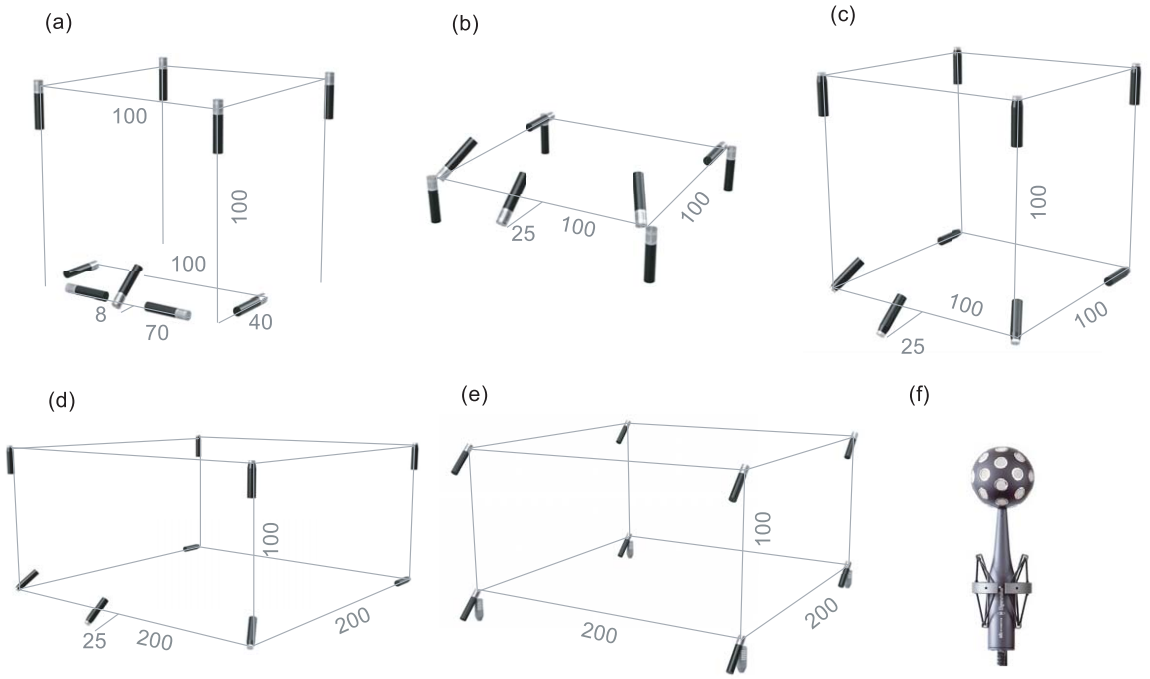

Fig. 1. Microphone arrays used for the recording and objective measurements: (a) OCT-3D, (b) PCMA-3D, (c) 2L-Cube, (d) Decca Cuboid, (e) Hamasaki Cube, (f) Eigenmike EM32. Unit for the numbers is cm. All microphones except for the Eigenmike EM32 and Hamasaki Square (Schoeps CCM8) were of the DPA d:dicate series. Detailed information on the polar patterns and microphone angles for each array can be found in [27].

nel time difference (ICTD) for imaging. An omni microphone typically has a better low-frequency extension than a directional microphone, which is why it is often more preferred to directional microphones by recording engineers. The exact microphone positions of the 2L-Cube are unclear from the available reference. In the 3D-MARCo session, the physical configuration of the base layer of 2L-Cube was identical to that of PCMA-3D (see Fig. 1). This allows a direct comparison between cardioid and omni polar patterns in an identical physical configuration. Furthermore, the omni polar pattern of the height layer microphones can be compared directly against the supercardioid of OCT-3D, which also has a 1 m x 1 m height layer at 1-m vertical spacing.

1.1.4 Decca Cuboid

The Decca Tree technique is widely used for large-scale orchestral recordings (it is a de-facto standard for film scoring). It employs three widely spaced omni microphones (FL-FR = 2 m to 2.5 m, FC-base = 1 m to 1.5 m), thus heavily relying on ICTD for phantom imaging. In 3D-MARCo, the traditional Decca Tree was augmented with rear microphones placed at 2 m behind the base point and height microphones 1 m above the base layer, thus named 'Decca Cuboid' here. The horizontal dimensions of this array are twice as large as 2L-Cube while keeping the vertical dimension the same. Therefore a greater amount of interchannel decorrelation can be expected. The FC microphone was placed 0.25 m in front of the base point instead of the originally used 1 m. The rationale for this was twofold: to be consistent with PCMA-3D and 2L-Cube for the comparison of the effects of different FL-FR spacings and avoid too strong a center image.

1.1.5 Hamasaki Square With Height

Hamasaki Square [33] is a popular technique for recording four-channel diffuse ambience. It was vertically extended based on Hamasaki and Baelen's approach [10]. The base layer consisted of four sideward-facing figure-of-eight microphones arranged in a 2 m x 2 m square. The height layer employed four cardioid microphones at two vertical positions from the base layer for a comparison purpose: 0 m (i.e., vertically coincident based on [8]) and 1 m (adapted from [10]). The original proposal by Hamasaki and Baelen [10] uses upward-facing supercardioids for the height channels. However, in 3D-MARCo, cardioid microphones were used instead and they faced directly away from the stage. This was considered to be more effective for suppressing direct sounds than using upward-facing supercardioids, particularly for the 0-m height layer.

1.1.6 Eigenmike EM32

Eigenmike EM32 by mhAcoustics is a spherical microphone array consisting of 32 omni capsules mounted on a small sphere. It can produce spherical harmonics with orders 1 to 4 for Ambisonic reproduction. In the current study, the first and fourth order Ambisonic reproductions were compared. Although an ideal Ambisonic reproduction requires a loudspeaker array configured in a regular polygon or polyhedral layout [34], it is possible to decode an Ambisonic recording to loudspeakers in an irregular arrangement (e.g., commercial 3D loudspeaker formats such as Dolby Atmos and Auro-3D as well as those recommended in ITU-R BS. 2051-2 [28]), using decoders optimized for the purpose (e.g., ALLRAD [35] and EPAD [36]).

Although the main focus of the current investigation is on perceptually motivated microphone arrays that were developed for an ITU-R-based nine-channel loudspeaker reproduction, Eigenmike EM32 was also included in the objective measurements for interested readers, as in practice Ambisonic recordings might be reproduced over such an irregular loudspeaker array more frequently than an ideal regular array. It is important to note that the results presented in this paper are specific to the loudspeaker configuration and decoder used and should not be generalized in terms of the performance of Eigenmike EM32 or Ambisonics. A separate investigation is required on the performance of Ambisonic decoding with different loudspeaker configurations.

1.2 Multichannel Room Impulse Responses

The 3D-MARCo database includes multichannel room impulse responses (MRIRs) captured in the St. Paul's concert hall (16 m x 30 m x 13 m; avg. RT = 2.1 s) in Huddersfield, UK, for thirteen source positions from -90° to 90° with 15° intervals (Fig. 2) using all of the microphone arrays described above. For the present study, the intermediate source position $+45^\circ$ was used for the computations of various parameters described in the following sections. This position was considered to be suitable for the purpose of this analysis as it would produce sufficient interchannel and interaural differences among the microphone signals, which would be necessary for observing differences among the microphone arrays in terms of localization and spatial impression.

For the acquisition of the MRIRs, the exponential sine sweep method [37] offered by the HAART software [38] was used. Genelec 8331A co-axial loudspeakers were used as sound sources. Their acoustic center was at 1.14 m above the floor.

2 OBJECTIVE MEASUREMENTS

A set of objective parameters measured in this study are listed below.

- The Interchannel level difference (ICLD) and interchannel time difference (ICTD) of interchannel crosstalk (ICXT).
- Temporal fluctuations of interaural level and time differences (ILD and ITD).
- Ear-signal's spectral distortion resulting from the ICXT of the height microphone layer.
- Interchannel correlation coefficient (ICC).
- Interaural cross-correlation coefficient (IACC).
- Direct-to-reverberant energy ratio (DRR).

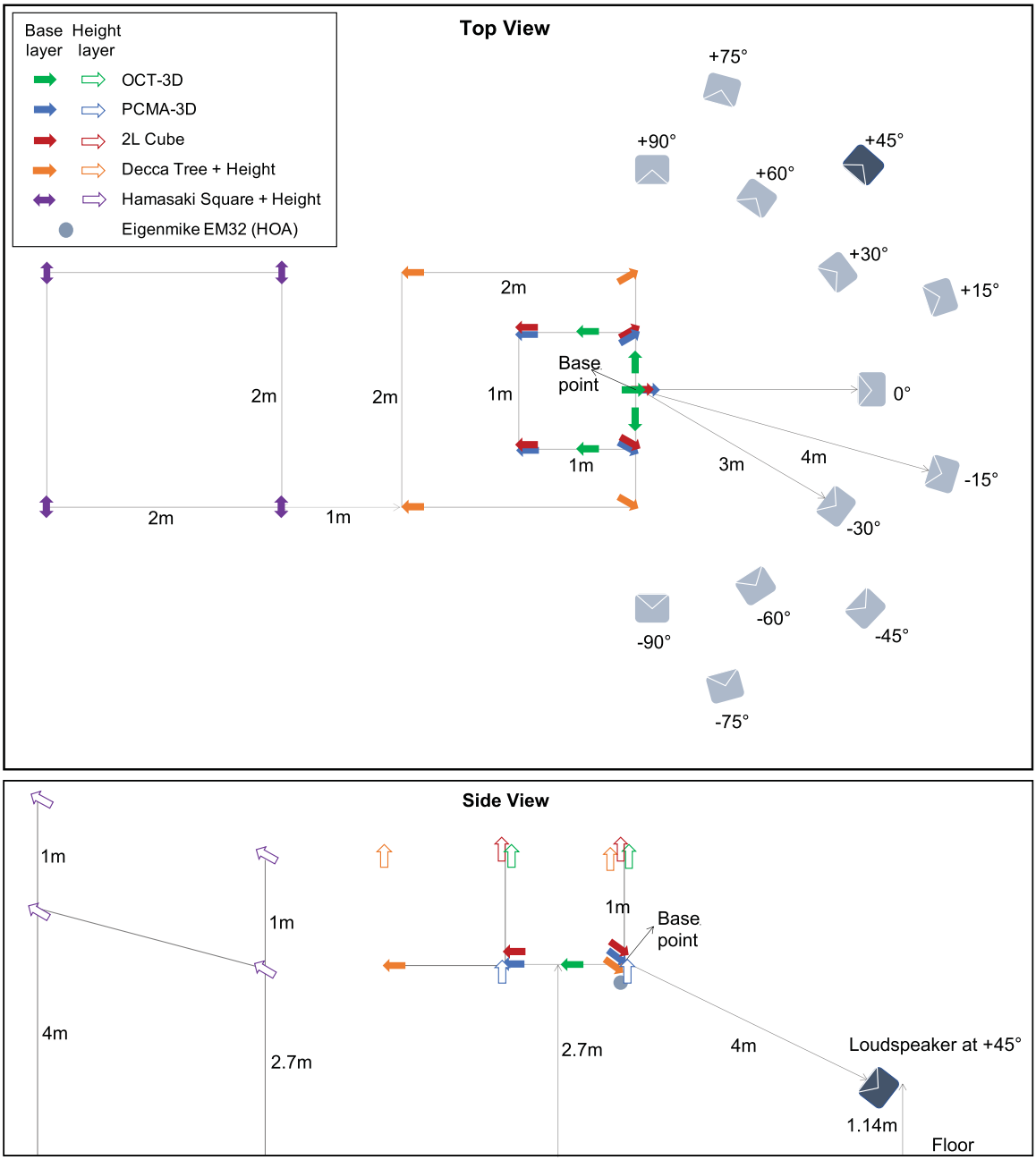

Fig. 2. Physical layout of the microphones and loudspeakers used for capturing the multichannel room impulse responses (MRIRs) in 3D-MARCo. For the objective measurements carried out in the present study, the MRIRs for the source at +45° were used.

These parameters were chosen because they were considered to be predictors for different types of perceptual attributes, such as horizontal and vertical image stability, tonal coloration, apparent source width (ASW), listener envelopment (LEV), vertical image spread, and perceived source distance. This section first describes the general methods employed for the measurements and details each of the parameters.

2.1 Methods

The analysis strategy used here was adapted from [8]; two types of signals were used for the analysis: (i) multichannel room impulse responses (MRIRs) taken directly from the database and (ii) binaural impulse responses from reproduction (BIRR), which were synthesized by convolving the MRIRs with the head-related impulse responses (HRIRs) for their corresponding loudspeaker positions from Table 1, thus creating ear-input signals from a virtual multichannel loudspeaker playback. The MRIRs were used for computing ICLD, ICTD, ICC, and DRR, whereas the BIRRs were used for measuring ILD, ITD, IACC, and the frequency spectra of ear-input signals. The use of room impulse responses for the current study allows for investigating source-related and environment-related perceptual properties of different microphone techniques. As commonly used in concert hall and room acoustics research,

the room impulse responses were segmented into the time windows of direct sound, early reflections, and reverberation, as required for the measured parameter.

Fig. 3 describes the overall workflow. The MRIRs of each spaced array was discretely routed to their corresponding loudspeakers from Table 1 (e.g., the front left and rear right microphone signals of an array to FL and RR, respectively). On the other hand, the raw signals of Eigenmike EM32 needed to go through a series of processing to obtain the loudspeaker signals. They were first converted into spherical harmonics using the EigenUnit plugin [39] that were then decoded to the loudspeakers configured as in Table 1. The ALLRADecoder plugin in the IEM plugin suite [40] was used since the ALLRAD method [35] was specifically designed for decoding an Ambisonic recording to irregular loudspeaker arrays such as the one used here (i.e., Table 1) and it is arguably the most widely used decoder for such a purpose.

The decoder weighting option in the plugin was set to 'basic,' which is optimized for an ITD synthesis in reproduction at frequencies below around 700 Hz [34]. From the authors' own subjective comparisons, the basic weighting produced more spacious and natural sound field than the 'max rE' or 'in-phase' weighting, which is optimized for ILDs at higher frequencies. Note that the measurement results to be presented in Sec. 3 are specific to the basic weighting and might be slightly different if the decoder used the max rE weighting or a dual-band approach where the basic and max rE weightings are used for lower and higher frequencies, respectively. It was not the scope of the present study to formally compare the performances of different types of decoders with different loudspeaker arrays. Readers who are interested in exploring various decoding options are recommended to use the IEM [40] or SPARTA [41] plugin suite on the Reaper session template provided with the 3D-MARCo database [27].

The loudspeaker signals were either kept as broadband or split into different frequency bands, depending on the parameters measured. The BIRRs were synthesized by convolving the MRIRs with the KU100 head-related impulse responses (HRIRs) taken from the SADIE II database [42]. The MRIRs or BIRRs underwent time-window segmentation as required for each of the parameters. Detailed descriptions for the segmentation are provided in each subsection below.

2.2 Parameters

2.2.1 Interchannel Level and Time Differences of Interchannel Crosstalk

In the context of microphone array design, interchannel crosstalk (ICXT) is defined as a direct sound captured by other microphones than the ones that are responsible for the localization of phantom image. Research suggests that a horizontal ICXT is significantly associated with perceptual effects such as locatedness (i.e., ease of localization) and source image spread [43], whereas an ICXT present in the height microphone signal (e.g., FLh) would cause the phantom source to be shifted upward unless it is suppressed by at least 7 dB in reference to the direct sound in the base microphone signal (e.g., FL) [32].

In the current study, the direct sounds picked up by other microphones than FL and FC microphones, which are primarily responsible for source imaging, are regarded as ICXTs. Therefore, the ICTD and ICLD of each signal are the properties of ICXT, which would influence the perceived characteristics of the resulting source image as the literature suggests [32, 43]. Here, the ICTD and ICLD of each signal to FL was calculated since the FL microphone was closest to the sound source at +45°, thus producing the earliest-arriving signal with the highest level among all microphones. The ICTDs were calculated as the lag (in ms) of the maximum absolute value of the normalized cross-correlation function (NCF) (Eq. (1)), using the MRIRs.

$$NCF_{t1,t2}(\tau) = \frac{\int_{t1}^{t2} x_1(t) \cdot x_2(t+\tau)\,dt}{\sqrt{\int_{t1}^{t2} x_1^2(t) \cdot \int_{t1}^{t2} x_2^2(t)\,dt}} \tag{1}$$

where x_1 and x_2 are channel signals, t_1 and t_2 are the lower and upper boundaries of time segment, and τ is the time lag. The time segment used for the computation was set to be long enough to include the direct sounds (i.e., impulses) of all microphones for each array ($t_1 = 0$ ms and $t_2 = 10$ ms). The lag limit was the same as the value of t_2. The ICLD of each signal compared to FL was computed as the energy difference between the signals in decibels.

2.2.2 Spectral Influence of ICXT

Tonal quality is often not discussed as much as spatial quality when discussing 3D sound recording and reproduction. However, it should be noted that the use of more channels presenting an ICXT has a potential risk of introducing a greater degree of spectral distortion in the ear-input signal due to the comb-filter effect, thus potentially influencing the perceived tonal characteristics of source images. To investigate this objectively, the difference of the magnitude spectrum of the left-ear input signal resulting from the combination of the base and height layers to that from the base layer only (i.e., delta spectrum) was measured. For this, the BIRRs up to 10 ms after the earliest direct sound were used. This was to include direct sounds present in all of the microphone signals and make the analysis window consistent across all of the arrays; the maximum ICTD to FL observed among all arrays was 9.5 ms for RRh-FL of Decca Cuboid (Fig. 4(b)).

2.2.3 Temporal Fluctuations of Interaural Level and Time Differences

ICLDs and ICTDs among the microphone signals are eventually translated into interaural level and time differences (ILD and ITD) at the ears in reproduction. It is well known that the ILD and ITD cues determine the perceived horizontal position of a sound image. However, when there is a modulation between two or more signals, the ILD and ITD tend to vary over time, and this has been found to be related to the perceived movement or spread the image depending on the fluctuation rate (i.e., the "localization lag"

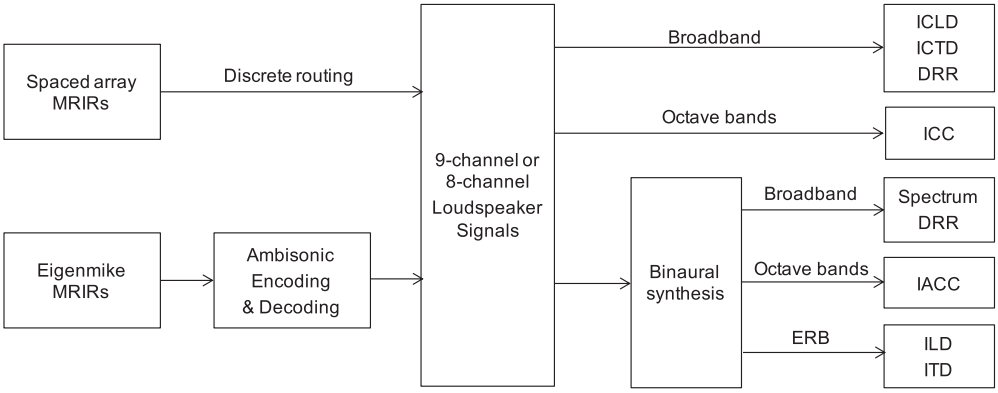

Fig. 3. Overall workflow for the objective measurements conducted.

phenomenon [44]). That is, at low rates of fluctuations (up to 3–20 Hz, depending on the experimental method and type of signal [44–46]), the image would be perceived to be moving between left and right, whereas higher rates would produce a stationary image with a spread (i.e., ASW). Based on this, measuring the fluctuations of ILD and ITD resulting from the reproduction of 3D microphone array signals would provide useful insight into the horizontal imaging stability and ASW.

To create a stimulus for measuring ILD and ITD fluctuations over time, for the +45° source position, the BIRRs up to 10 ms after the earliest direct sound were first convolved with a 10-second-long pink noise signal (20 Hz to 20 kHz) and an anechoically made trumpet recording from [47] for each array. As mentioned in the previous section, the 10-ms analysis window of the BIRR included the direct sounds captured by all microphones for each array. The trumpet recording was chosen as it has time-varying musical notes, whereas the noise is broadband and time-consistent.

The convolved stimuli were split into 64 equivalent rectangular bands (ERBs) through a Gammatone filter bank [48]. Half-wave rectification and a first-order low-pass filtering at 1 kHz were applied to mimic the breakdown of the phase-locking mechanism as used in [49, 50]. The resulting signals were then time-segmented into 50%-overlapping Hann-windowed 50-ms frames. The ITD (time delay of the left ear signal to the right one) was computed as the lag (in ms) of the maximum absolute value of the NCF (Eq. (1)) with the lag limit of 1 ms [51]. The ILD was computed as the energy difference of the left ear signal to the right one in decibel. Then, for each frame, the ITDs were averaged for the ERBs with the center frequency of 1.47 kHz and below, whereas the ILDs were averaged for the ERBs with center frequencies from 1.62 kHz to 19 kHz.

2.2.4 Interchannel Correlation Coefficient (ICC)

The magnitude of interchannel correlation is associated with auditory image spread in horizontal stereophonic re-

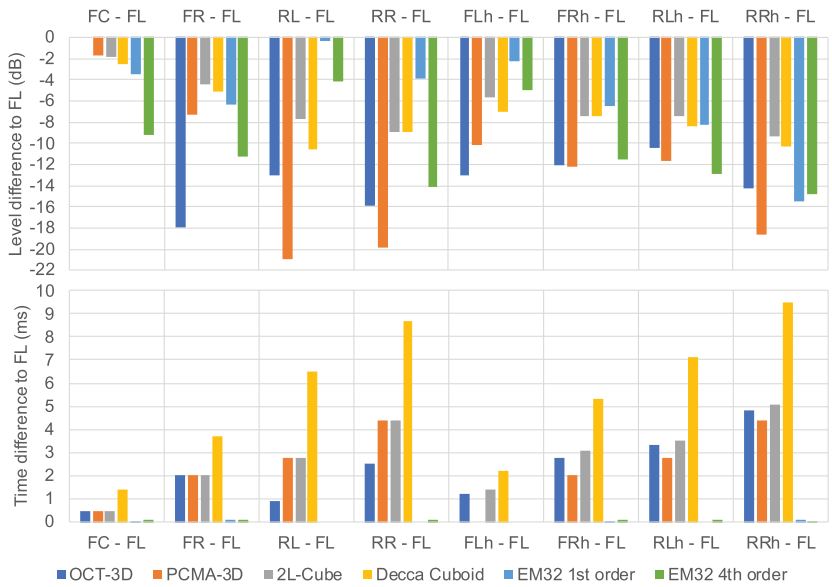

Fig. 4. Interchannel level and time differences (ICLD and ICTD) of each microphone to FL, measured using the energy of the direct sound portion (0–2.5 ms) of the impulse responses captured for the +45° source position. ICLDs were not calculated for the Hamasaki Square arrays as their aim is to capture ambience.

production as well as listener envelopment (LEV) [33, 52, 53]. It is also related to the size of listening area (i.e., the more decorrelated, the wider the sweet spot) [33]. For the present investigation, interchannel correlation coefficients (ICCs) were calculated as the absolute value of the NCF (Eq. (1)) with $\tau = 0$. Cross-correlation as in SEC 2.2.1 was not used since the motivation here was to investigate the magnitude of differences between the fixed microphone positions rather than finding the ICTD. As with ICXT, ICC was computed for each of the microphone signals against FL. Additionally, ICCs for the symmetrically arranged microphone pairs RL-RR, FLh-FRh, and RLh-RRh were measured.

For computing the ICCs, the MRIRs were first split into nine octave bands with their center frequencies ranging from 63 Hz to 16 kHz, using an eighth-order biquad linear-phase filter (48 dB/oct). Then each band signal was segmented into early and late portions (i.e., ICC Early (E): $t_1 = 0$ ms to $t_2 = 80$ ms; ICC Late (L): $t_1 = 80$ ms to $t_2 = 2{,}100$ ms) in order to predict differences in source-related and environment-related spatial attributes. The 80-ms boundary point between the two segments is typically used for musical sources in concert hall research [54]. ICC was calculated for each octave band, after which the results were averaged for low (63 Hz, 125 Hz, and 250 Hz), middle (500 Hz, 1 kHz, and 2 kHz), and high (4 kHz, 8 kHz, and 16 kHz) bands. Here the results are referred to as ICC E(or L)$_{Low}$, ICC E(or L)$_{Mid}$, ICC E(or L)$_{High}$.

2.2.5 Interaural Cross-Correlation Coefficient (IACC)

IACC is widely known as a parameter to predict the perceived horizontal width of an auditory image. It is defined as the maximum absolute value of the NCF (Eq. (1)) obtained over the lag range of 1 ms and +1 ms of the ear-input signals. Hidaka et al. [54] found that ASW and LEV in concert halls were best predicted using the average of the IACCs for the octave bands with the center frequencies of 500 Hz, 1 kHz, and 2 kHz, proposing objective measures IACC E3 for ASW and IACC L3 for LEV. IACC E3 is measured for the early time segment of binaural room impulse responses ($t_1 = 0$ ms to $t_2 = 80$ ms), while IACC L3 is computed for the late segment ($t_1 = 80$ ms to $t_2 = 750$ ms). For the current measurement, IACC E3 and IACC L3 were computed using BIRRs synthesized for each of the base and height loudspeaker layers separately as well as both layers. This was to demonstrate the predicted subjective effects of adding the height layer to the base layer in terms of ASW and LEV.

2.2.6 Direct-to-Reverberant Energy Ratio (DRR)

The direct-to-reverberant energy ratio (DRR) is widely known as an absolute measure for perceived auditory distance in rooms [55, 56]. It is typically measured using a BRIR captured using an omni-directional microphone. In the context of microphone array recording, the DRRs of ear-input signals resulting from multichannel reproduction as well as those of individual microphone signals might be a useful indicator for the perceived distance of a phantom image. The integration time window used for the direct sound energy was 2.5 ms since it is approximately the duration of anechoic HRIR and short enough to exclude the first reflection [55]. For the DRRs of the ear-input signals, however, it would be necessary to include the direct sounds from all of the microphone signals for each array. Therefore, the time window was determined by 2.5 ms plus the maximum ICTD from the earliest signal (FL in the current case).

3 RESULTS AND DISCUSSIONS

3.1 Level and Time Differences of Interchannel Crosstalk

Fig. 4 shows the level and time differences of each channel signal to the FL signal, calculated for the direct sound portion of each signal (up to 2.5 ms after the initial impulse). As mentioned in SEC. 2.2.1, FL is used as a reference here since it is the microphone closest to the sound source used in this analysis (45° to the left from the centre). Based on [29], FL and FC are mainly responsible for source imaging and all other microphone signals are assumed to be ICXT in this case. Hamasaki Square was excluded for this analysis since it is designed for mainly capturing ambience rather than direct sound.

Looking at the horizontal channel pairs first, it can be observed that OCT-3D had a substantially weaker ICXT (−18 dB) than all other arrays for FR-FL. This was expected as the front triplet of OCT-3D is specifically designed to reduce ICXT by using sideward-facing supercardioids as described in Sec. 1.1. However, for the rear microphones RL and RR, it can be seen that PCMA-3D suppressed the ICXT more effectively than OCT-3D for the given source position. Looking at the ICTD, the RL of PCMA-3D was 2.8 ms delayed to FL, whereas that of OCT-3D was delayed by 0.9 ms. From these observations, the following can be suggested. OCT-3D would likely have a better locatedness than PCMA-3D for frontal phantom images due to the stronger suppression of ICXT, whereas the latter would produce a larger ASW. Although the ICTD between the front and rear channels, for both OCT-3D and PCMA-3D, is large enough to trigger the precedence effect [57] in combination with the ICLD, the better front-rear separation of PCMA-3D might provide more headroom for increasing the level of the rear ambience without affecting the accuracy of frontal localization.

2L-Cube and Decca Cuboid generally had stronger ICXT than OCT-3D and PCMA-3D due to the use of omni-directional microphones. The ICTDs of all channels to FL were larger than 1 ms for all pairs, which would be sufficient to trigger the precedence effect for localization between the horizontal channels. However, as reported in [31], the precedence effect would not operate between vertically oriented loudspeakers by ICTD alone; at least a reduction of 7 dB would be required to avoid the localization uncertainty [32]. 2L-Cube and Decca Cuboid in the current recording setup produced the ICXT reduction of 5.7 dB and 7 dB

for FLh, respectively. This is close to the 7 dB threshold but considerably smaller compared to OCT-3D (13 dB) and PCMA-3D (10 dB). Based on this, it can be suggested that the height channels of OCT-3D and PCMA-3D could be boosted by around 6 dB to 3 dB, respectively, without affecting the localization of the source image, whereas doing the same with 2L-Cube or Decca Cuboid would not only cause a loudness increase but also shift the image upward.

The Eigenmike conditions generally show that the fourth order rendering had a considerably lower level of ICXT than the first order rendering, which was an expected result due to the increased spatial resolution of the higher-order Ambisonics. The channel separation of the first order was found to be particularly small for RL-FL (−0.3 dB) and FLh-FL (−2.3 dB). It is important to note, however, that in Ambisonic decoding, all loudspeaker signals contribute to the synthesis of binaural cues for sounds arriving from different directions. Therefore, the small amount of level difference between specific channels does not directly indicate that the accuracy of imaging would be poor. However, the small channel separation would likely cause unstable phantom imaging outside the small sweet spot [58].

3.2 Spectral Influence of Interchannel Crosstalk

The results for the spectral magnitude measurements are shown in Fig. 5. The delta plots in the right columns represent the effect of adding the height layer to the base layer in terms of the ear-input signal spectrum. A positive value in the plots indicates that the height layer signals were added to the main layer signals constructively at the ear, whereas a negative value means that the addition of the height layer signals was spectrally destructive to the ear input signals of the base layer.

The results generally show that the height layer of the vertically spaced arrays had a noticeably stronger spectral influence on the ear signal than that of the vertically coincident arrays. As can be observed from the delta plots in Fig. 5(b), the main and height layers of PCMA-3D were summed at the ear constructively at almost all frequencies up to about 8 kHz with only a few erratic peaks, whereas the height layers of 2L-Cube and Decca Cuboid produced substantial amount of magnitude fluctuation depending on the frequency. OCT-3D also had a similar pattern but the magnitude and frequency of the peaks and dips were smaller compared to 2L-Cube and Decca Cuboid.

These results can be explained as follows. As shown in previous section, the height layer signals of 2L-Cube and Decca Cuboid, which use omni microphones, generally had a higher level of ICXT than those of OCT-3D and PCMA-3D using upward-facing supercardioids. Furthermore, the main and height layers of the latter arrays were vertically spaced, producing ICTDs between the vertical microphones, e.g., FL-FLh. Consequently, when all of the signals are summed at the ear, 2L-Cube and Decca Cuboid would suffer from a stronger comb-filter effect than the other arrays with weaker ICXTs.

The height layer of the coincident array Eigenmike had the minimal spectral effect, producing only increase in level up to about 8 kHz. This was expected as the ICTDs were zero or negligibly small as shown in Fig. 5. However it should be noted that, unlike the perceptually motivated arrays that treat the base and height layers separately for source and environmental sound imaging, Ambisonic decoding requires all of the signals from both layers to be presented for the reconstruction of sound field. Therefore the delta spectra for the Eigenmike conditions do not represent a tonal coloration of the source image caused by the height layer but rather the spectral contribution of the height layer on the complete construction of the source image.

The above results imply potentially substantial differences among the arrays in perceived tonal color. However, the subjective interpretation of tonal color seems to be a complex cognitive process, which may depend on not only the type of sound source but also one's experience and expectation. Theile's 'association model' [59] suggests that the perception of the tonal color of a phantom image is also related to localization; the audibility of tone coloration depends on the magnitude of spectral distortion against a reference ear signal spectrum associated with the perceived direction of a certain phantom image. Based on this, it may be that the spectral differences observed in the current analyses would be most audible for a single source but less so for complex ensemble sources. This will be confirmed in subjective studies to follow in the future.

3.3 ILD and ITD Fluctuations Over Time

Fig. 6 shows the time-varying ILDs and ITDs measured for the binaural signals resulting from the reproduction of the multichannel room impulse responses convolved with pink noise and anechoic trumpet sources. To quantify the magnitude of fluctuation, three standard deviations (3SD) are presented in Table 3. For the noise, differences among the arrays in the 3SD of ILD was minimal (< 0.37 dB). However, those in ITD were considerably large, with 2L-Cube having the highest value of 3SD (0.52 ms), followed by Decca Cuboid, PCMA-3D, OCT-3D, and the Ambisonic conditions. This generally suggests that the spaced 3D microphone techniques cause a greater magnitude of ITD fluctuation over time than the coincident techniques, which is also in line with Lipshitz's [60] observation on two-channel stereo microphone techniques.

The differences in ITD fluctuation observed for the noise source seem to be related to ASW perception rather than image movement since the fluctuation was constantly random and rapid for all arrays. It is not possible to derive an exact fluctuation rate in the same controlled way as in studies using pulse train or modulated noise (e.g., [45, 44]). Instead the number of flips in the motion of ILD and ITD was counted for each array. The rate of ILD flip was between 19 Hz and 21 Hz, whereas the ITD flip rate was between 21 Hz and 31 Hz, which are considered to be high enough to suggest an ASW perception based on [44, 45].

For the trumpet, on the other hand, a large degree of image movement in accordance with the time-varying note of the performance could be anticipated from the plots in Fig. 6, depending on the type of microphone array. For

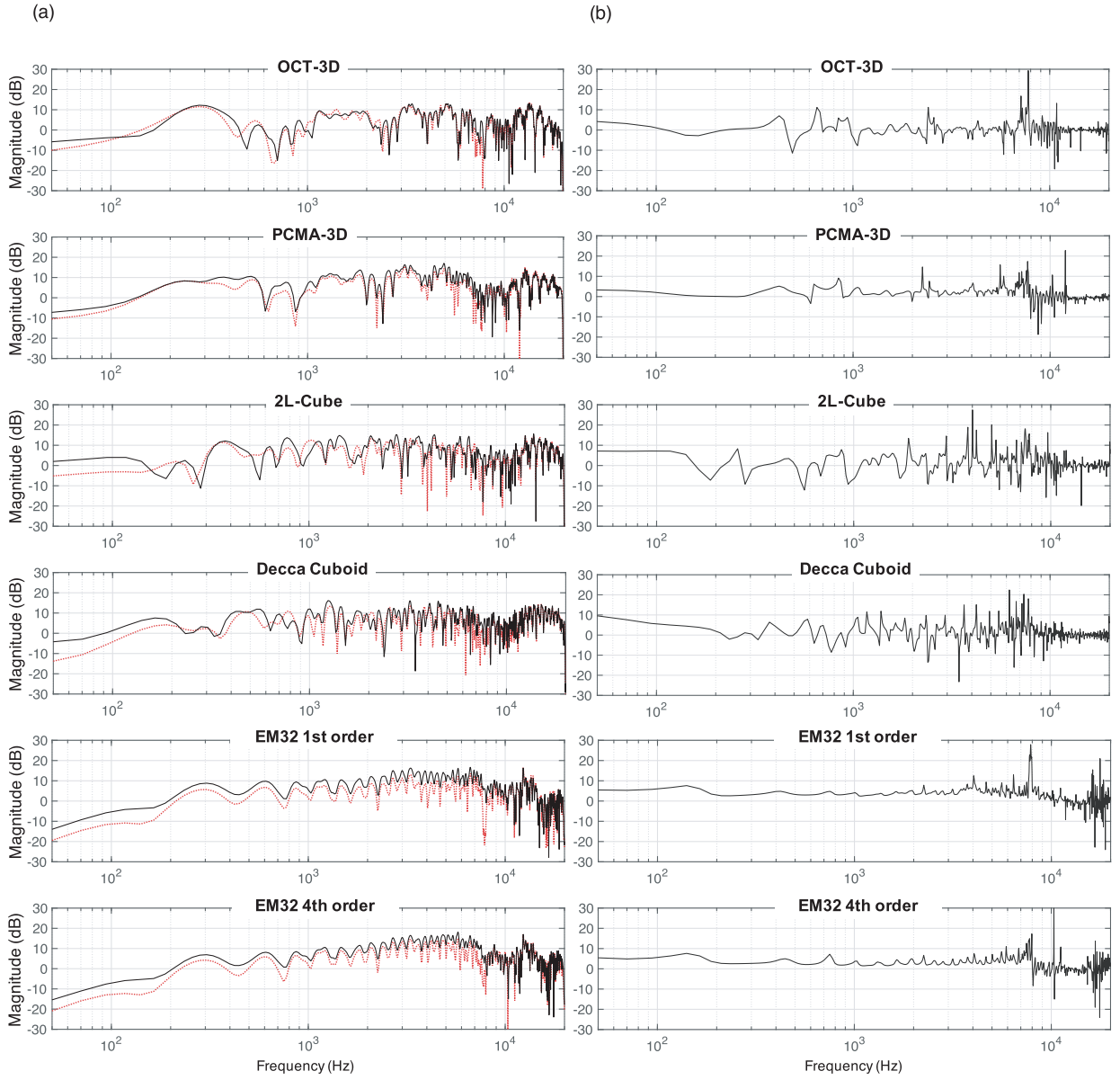

Fig. 5. Spectral magnitudes of the left-ear input signal of the binaural impulse responses resulting from the loudspeaker playback of multichannel impulse responses. (a) Measurements for both the base and height layers (solid lines) and those for the base layer only (dotted lines); (b) difference of both layers to the base-layer-only in the spectral magnitude (i.e., the spectral effect of the height layer).

Table 3. Means and three standard deviations (3SDs) of interaural ITD and ILD fluctuations over time.

	Noise				Trumpet			
	ILD (dB)		ITD (ms)		ILD (dB)		ITD (ms)	
Array	Mean	3SD	Mean	3SD	Mean	3SD	Mean	3SD
OCT-3D	2.11	0.57	–0.21	0.17	3.88	5.15	–0.05	0.57
PCMA-3D	1.56	0.71	–0.28	0.24	2.88	4.93	–0.12	0.32
2L-Cube	0.83	0.91	–0.33	0.52	1.29	9.56	0.11	0.61
Decca Cuboid	1.22	0.85	–0.16	0.43	1.31	6.55	–0.14	0.65
EM32/ALLRAD 1st	6.91	0.55	–0.36	0.13	7.98	2.64	–0.22	0.14
EM32/ALLRAD 4th	5.55	0.54	–0.34	0.15	7.28	2.28	–0.24	0.33

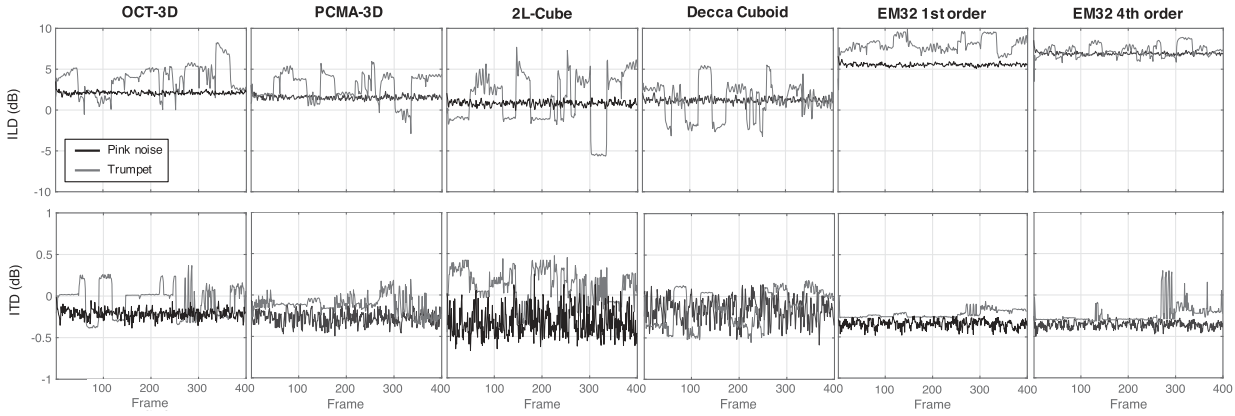

Fig. 6. ILDs and ITDs measured for the 50%-overlapping 50-ms Hann-windowed frames of 10-second-long pink noise (black) and anechoic trumpet (grey). The ILD and ITD for each frame are the averages of ILDs and ITDs computed for the ERBs with the center frequencies between 1.62 kHz and 19 kHz and for those up to the center frequency of 1.47 kHz, respectively.

OCT-3D, 2L-Cube, and Decca Cuboid, which are in the horizontally and vertically spaced (HVS) category of 3D microphone arrays, the ILDs and ITDs had large occasional shifts between positive and negative values. PCMA-3D, which is a horizontally spaced and vertically coincident (HSVC) array, had a moderate ITD fluctuation pattern, with a smaller 3SD than the HVS arrays for both ILD and ITD. The Ambisonic arrays had the most consistent ILDs and ITDs among all arrays, with the smallest 3SDs for ILD and ITD as can be observed in Table 3. This seems to indicate that a larger ICTD between microphone signals would lead to a greater degree of ILD and ITD fluctuations for musical signals with time-varying single notes and thus a poorer imaging stability.

3.4 Interchannel Correlation Coefficient (ICC)

Fig. 7 presents the results of the ICC analyses. At a glance, it is apparent that the low band ICCs were generally higher than the middle and high band ones in both segments for all spaced microphone arrays, with the high band values being close to 0. The difference between the early and late segments was also minimal for most spaced array conditions. On the other hand, the ICCs for the Eigenmike conditions were generally higher than those for the spaced arrays, regardless of the bands, as one might expect from a coincident array.

Differences between the spaced microphone arrays appear to be most obvious at the low bands. For FL-FR, the Decca Cuboid had the lowest ICC E_{Low} (0.19), which was expected due to the largest microphone spacing (2 m) and the resulting ICTD of 3.7 ms (Fig. 6(b)). However, OCT-3D had a considerably lower ICC E_{Low} (0.33) than PCMA-3D (0.53) and 2L-Cube (0.52) even though they all had the same ICTD of 2 ms (Fig. 6(b)). This seems to be associated with the use of the $\pm 90^\circ$-facing supercardioid microphones for OCT-3D. That is, FR not only suffered less from ICXT as discussed earlier (Fig. 6) but also would have captured strong early reflections predominantly from the right-hand side while suppressing those from the left-hand side, which would eventually have lowered the ICC.

Observing FL-FLh, PCMA-3D had substantially higher IACC E and IACC L than OCT-3D, 2L-Cube, and Decca Cuboid across all of the frequency bands. This is likely to be due to the vertically coincident configuration of the microphones. On the other hand, the other vertical pairs of PCMA-3D (FL-FRh, FL-RLh, and FL-RRh) still had at least 1-m spacing between the microphones and therefore their ICCs were comparable to those of the other spaced main arrays in general. Gribben and Lee [61, 62] found that in a nine-channel loudspeaker reproduction, the effect of vertical ICC on vertical image spread (VIS) was largely insignificant for low frequencies but significant for frequencies above about 1 kHz, albeit only slight. The current results show that the ICCs of the vertical pairs for all of the spaced arrays apart from PCMA-3D were very low (about 0.1 or below) for the middle and high frequency bands. Based on the above it is hypothesized that, if any differences in perceived VIS were perceived among the spaced main arrays, it would be due to ICXT rather than ICC.

Griesinger [53] claims that for reverberation in the rear channels, decorrelation at low frequencies would be particularly important for increasing the magnitude of listener envelopment (LEV). Looking at the ICC L_{Low} values for RL-RR in the current results, Decca Cuboid with the horizontal spacing of 2 m and Eigenmike first order had the lowest (0.19) and highest (0.63) values among all, respectively. The difference between PCMA-3D (0.36) and 2L-Cube (0.34), which share the same horizontal spacing of 1 m, was negligible, while OCT-3D with a 0.7 m horizontal spacing had a slightly higher ICC L (0.44) than them. A similar pattern was found for RLh-RRh. From these results, it could be predicted that the perceived magnitude of LEV would be correlated with the horizontal microphone spacing. However, as will be discussed further in SEC. 3.6, the ICC-based prediction by Griesinger seem to conflict with an IACC-based prediction based on Hidaka et al. [54].

For the Eigenmike conditions, it appears that the difference between the first and fourth orders generally became larger with an increasing frequency band, depending on the channel pair. For instance, the fourth order had a dra-

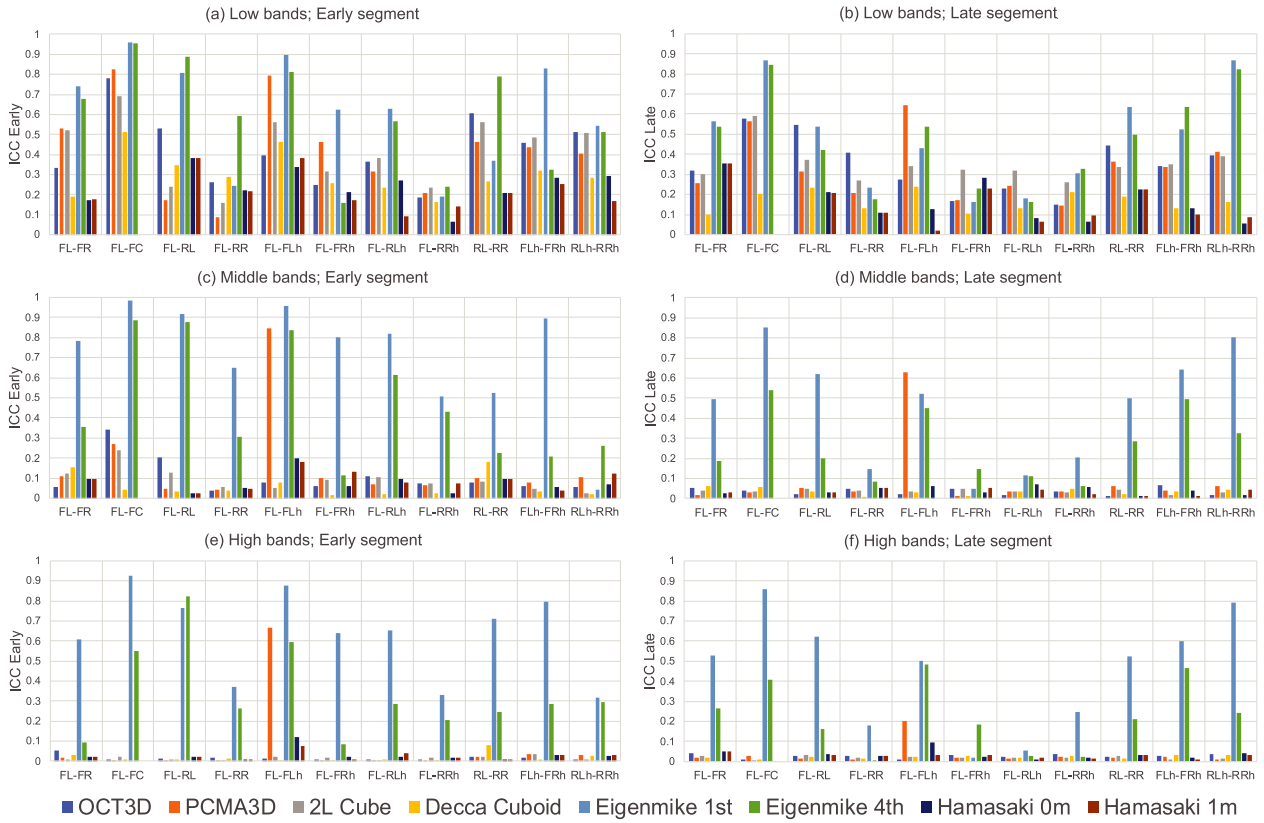

Fig. 7. Interchannel correlation coefficients (ICCs) for different pairs of microphone signals, computed using multichannel room impulse responses split into octave bands; average ICCs for low bands (centered at 63 Hz, 125 Hz, and 250 Hz), middle bands (500 Hz, 1 kHz, and 2 kHz), high bands (4 kHz, 8 kHz, and 16 kHz); segmentized into the early (0 to 80 ms) and late (80 to 2,100 ms) of the impulse responses.

matic decrease of ICC E from 0.67 to 0.1 for FL-FR as the band increased from low to high, while the first order only had a small change between 0.78 and 0.6. The ICCs for FL-RL, however, were consistently high (0.76–0.92) and had a minor difference between the first and fourth orders regardless of the frequency band. This may suggest that, in the current nine-channel reproduction using an irregular loudspeaker array in a controlled listening room, the well-known limitation of Ambisonic loudspeaker reproduction regarding phasiness would still exist during a front-back head movement even at the higher order.

Furthermore, it is worth noting that the ICCs of the Ambisonic loudspeaker signals would vary with different decoders. The ALLRA decoder (ALLRAD) used for the current analysis [38] was set to use the 'basic' weighting, which is optimized for an ITD synthesis in reproduction at frequencies below around 700 Hz [34]. The result might be different if the decoder used the 'max rE' weighting, which is optimized for ILDs at higher frequencies, or a dual band approach where the basic and max rE weightings are used for lower and higher frequencies, respectively.

3.5 Interaural Cross-Correlation Coefficient (IACC)

The results of IACC measurements are plotted in Figs. 8(a) to 8(d). Fig. 8(d) plots the differences of the IACCs for both layers to those for the base layer, which indicates the contribution of the height layer to the overall IACC. Firstly, for IACCs computed for both layers (Fig. 8(c)), the IACC E3 values for the Eigenmike conditions were substantially higher than those for the horizontally spaced arrays, following a trend similar to the ICC results. On the other hand, the differences among the arrays in IACC L3 appear to be negligible. This seems to suggest that the differences between the spaced and coincident arrays might be large in ASW but little in LEV.

However this conflicts with what the ICC L_{Low} measurement results suggest based on [53]: a greater degree of low frequency decorrelation would lead to a greater magnitude of LEV. Hidaka et al. [54] proposed the use of the octave bands centered at 500 Hz, 1,000 Hz, and 2,000 Hz for IACC since at lower frequency bands IACC values would be inherently high (close to 1) due to the ear-to-ear spacing being only around 17 cm. On the other hand, it is possible to have a low ICC between microphone signals (close to 0) at low frequencies depending on the spacing as shown in Fig. 7. However, when the signals are summed at the ears, the resulting IACC at those frequencies would still be high. Therefore, it is not yet clear which measure is more correlated with the actual perception of LEV. This will be answered from subjective listening tests to be conducted in the future.

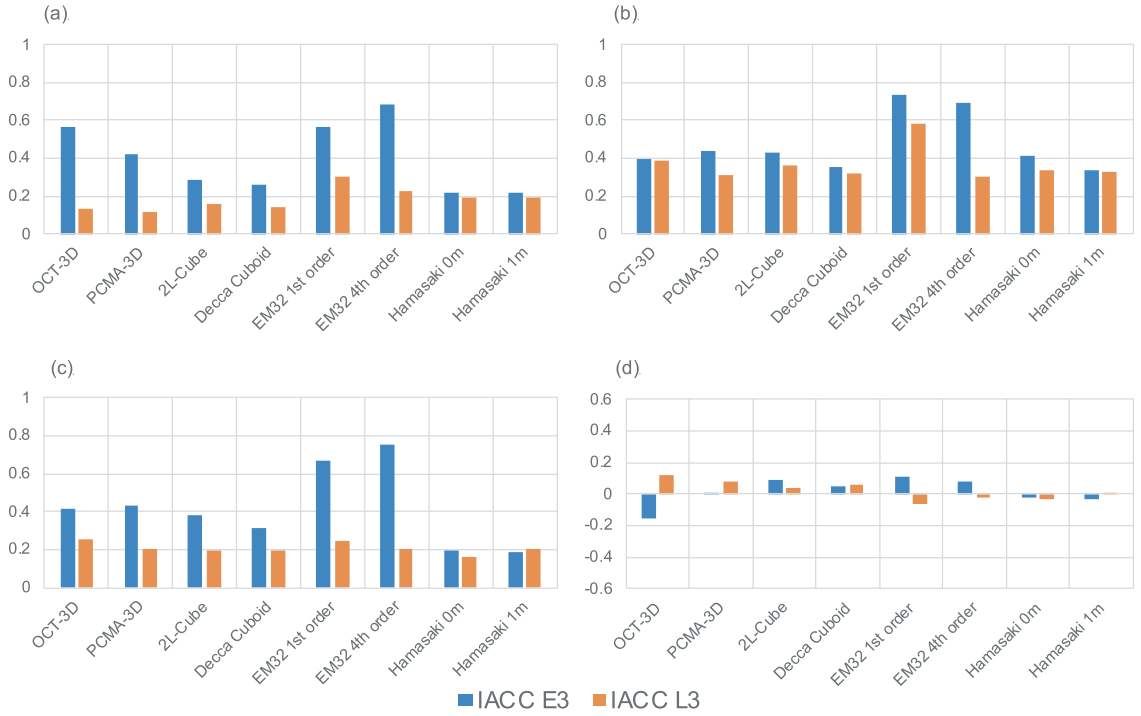

Fig. 8. Interaural cross-correlation coefficients (IACCs) for ear-input signals resulting from different microphone signals reproduced from a binaurally synthesized nine-channel 3D loudspeaker system. (a) IACC for the base layer only, (b) IACC for the height layer only, (c) IACC for both layers, (d) IACC difference (both layers base layer).

It is also interesting that, although IACC L3 for the height-layer-only condition was considerably higher than that for the base-layer-only in general, the influence of the height layer on the overall IACC L3 was minimal; the largest difference of both layers to the base-layer-only condition was 0.12 for OCT-3D (Fig. 8(d)). This seems to suggest that the height layer would not contribute to LEV in general.

It can also be observed that differences among the spaced main arrays in IACC E3 for the base layer appear to be greater than those for the height layer. However, with both layers presented, the differences become noticeably smaller, thus smaller differences in ASW. This is mainly due to the decrease in IACC E3 for OCT-3D (−0.15) and increase for 2L-Cube (+0.1) and Decca Cuboid (+0.05) when the height layer was added. Although these changes appear to be small, their effect on ASW may still be slightly audible since the just noticeable difference (JND) of ASW is known to be 0.075 [63]. PCMA-3D was hardly influenced by the height layer in IACC E3.

In addition, the two vertical spacings of 0 m and 1 m for the Hamasaki Square variants did not produce any meaningful differences in either IACC E3 or IACC L3. This suggests that there would be no benefit of raising the height layer of an ambience array above its base layer in terms of ASW and LEV. This complements the findings by Lee and Gribben [8], who showed that vertical spacing of a 3D main microphone array did not have a significant effect on perceived spatial impression.

3.6 Direct-to-Reverberant Energy Ratio (DRR)

Fig. 9 shows the DRR measurement results. At a glance it is obvious that the Hamasaki Square signals had the lowest DRRs as the microphone arrays aim to maximally suppress direct sounds and they were placed further away from the source; the negative values indicate that the direct sound energy was smaller than the reverberant energy. For the other arrays, differences among them varied depending on the channel. For the frontal channels in the main layer (FL, FC, and FR), most of the DRRs were positive and their differences varied within about 3 dB, but the OCT-3D's FR had substantially lower DRR (−8 dB) compared with the other spaced arrays (2.4–2.8 dB). This is related to the large amount of ICXT suppression achieved by the use of side-facing supercardioid microphone.

For RL and RR among the main microphone arrays, PCMA-3D had the lowest DRRs overall, followed by OCT-3D, owing to the use of backward-facing cardioids. The DRRs for 2L-Cube and Decca Cuboid are closer to 0, which is likely to be due to the use of omni-directional microphones. For the height channels, the DRR is the lowest with OCT-3D for all channels apart from RRh. It is noticeable that the DRRs for the Eigenmike conditions were mostly positive and substantially higher than the other arrays for all of the height channels as well as RL, regardless of the order.

However, looking at the DRRs of the ear-input signals from all of the individual channel signals, the maximum difference among the main arrays was relatively small: 2.4 dB between 2L-Cube and Eigenmike fourth for the left ear,

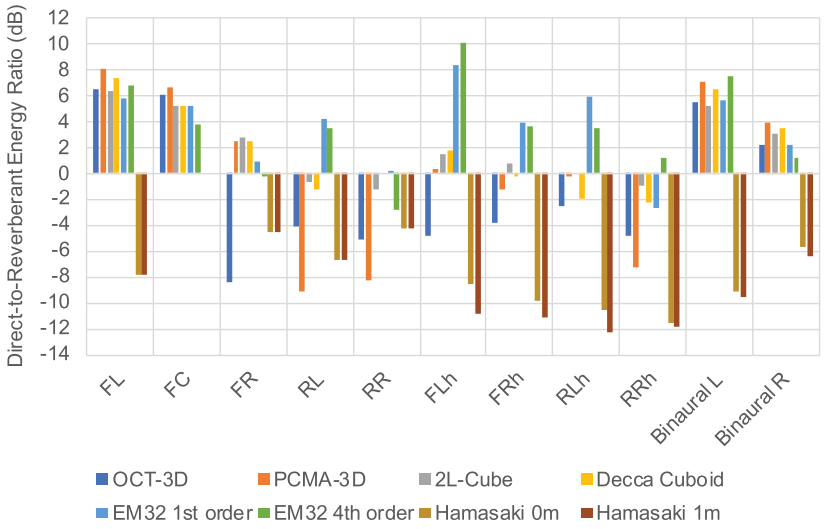

Fig. 9. Direct-to-Reverberant Ratio (DRR) for each microphone and ear-input signal.

and 2.7 dB between Eigenmike fourth and PCMA-3D for the right ear. The question of whether these differences are meaningful or not in terms of perceived source distance will be answered in a future subjective study using the recordings from the database. However, an insight could be gained from the literature on JND for DRR. Larsen et al. [64] reported that JNDs were 2–3 dB for the reference DRRs of 0 dB and 10 dB, and 6–9 dB for −10 dB and 20 dB DRRs, whereas Zahorik [55] found that the JNDs were consistently 5–6 dB for the reference DRRs of 0 dB, 10 dB, and 20 dB. This discrepancy might be due to different experimental conditions used in the studies. Whichever JND is trusted, it would seem that the maximum difference of 2.4–2.7 dB in DRR observed here alone suggests a small to no audible effect on perceived source distance. It is not clear yet whether it is the DRRs of individual channel signals or those of resulting binaural signals that would determine the perceived distance. This should be clarified in a future subjective study.

4 CONCLUSIONS

This paper presented the objective measurements for various types of 3D microphone arrays from the 3D-MARCo database, which is an extensive set of sound recordings of various musical performances and room impulse responses produced in a concert hall using various different 3D microphone arrays. The microphone arrays investigated in the present study were OCT-3D, PCMA-3D, 2L-Cube, Decca Cuboid, Eigenmike EM32, and Hamasaki Square with 0-m and 1-m vertical spacings of the height layer. Various objective parameters that might be associated with different perceptual attributes were computed, comprising the level and time differences to interchannel crosstalk (ICXT), the spectral influence of ICXT, fluctuations of interaural level and time differences (ILD and ITD) over time, interchannel correlation coefficient (ICC), interaural cross-correlation coefficient (IACC), and direct-to-reverberant energy ratio (DRR). The aim of these measurements was to produce hypotheses for future subjective studies to be conducted on the perceptual differences between the arrays. The observations from the results generally suggest the following.

There were substantial differences among the investigated microphone arrays in the amount of both horizontal and vertical ICXT, and this was found to be associated to the differences in the amount of spectral distortion in the ear signal as well as in the magnitudes of ILD and ITD fluctuations over time. From this, it is expected that the arrays would have audible differences in perceived timbral characteristics as well as the locatedness and spread of phantom image.

It is hypothesized that the arrays would have considerable differences in the perceived magnitudes of apparent source width (ASW) and the size of listening area due to the large differences in ICC between the early segments of the main layer impulse responses. Considerable differences in vertical ICC were also observed. However, based on previous research [65], this is hypothesized to have a minimal effect on perceived vertical image spread.

The analysis of IACC suggests that the addition of the height layer to the base layer in reproduction would have little effect on ASW and LEV regardless of the array type, even though the two layers might have audible differences in those attributes when they are reproduced independently. The differences between the microphone arrays in the DRRs of ear-input signals resulting from the virtual nine-channel loudspeaker reproduction were around or below the just noticeable difference of perceived auditory distance (i.e., 4 dB), even though the DRRs of individual microphone signals had considerably larger differences among the arrays. This raises a question of whether perceived source distance would be determined by the channel-dependent DRR or the DRR of the final ear signal.

Future studies will include the verbal elicitation of perceptual differences among the microphone arrays to establish a set of attribute scales that will then be used for grading

the microphone arrays. Subjective data resulting from the grading experiment will be compared against the objective measurements presented in this paper. From this, the perceptual weightings of the objective parameters will be determined to develop a statistical model for 3D acoustic recording quality evaluation.

5 ACKNOWLEDGMENT

This project was supported by Innovate UK (grant ref: 105175) and the University of Huddersfield (grant ref: URF 510-01). The authors would like to thank Eddy Brixen and DPA Microphones for providing the microphones used for the recordings, Paul Mortimer of Emerging UK and Claude Cellier of Merging Technologies for providing a Horus audio interface and AD8P microphone preamps for the recording, and Aki Mäkivirta and Siamäk Naghian of Genelec for providing the 8331A loudspeakers used for the impulse response acquisition. The authors are also grateful to Bogdan Bacila for designing and 3D-printing microphone mounts and tube joiners, and all of the members of the Applied Psychoacoustics Lab who assisted on the project.

6 REFERENCES

[1] Inc. Dolby Laboratories, "Dolby Atmos," https://www.dolby.com/technologies/dolby-atmos (accessed Oct. 26, 2021).

[2] Auro Technologies, "Auro-3D / Auro Technologies: Three-Dimensional Sound," https://www.auro-3d.com (accessed Oct. 26, 2021).

[3] DTS, Inc., "DTS:X - DTS," https://dts.com/dtsx/ (accessed Oct. 26, 2021).

[4] Sony Europe B.V., "360 Reality Audio," https://www.sony.co.uk/electronics/360-reality-audio (accessed Oct. 26, 2021).

[5] ITU-R, "Multichannel Sound Technology in Home and Broadcasting Applications," Report ITU-R BS.2159-8 (2019 Jul.).

[6] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, "MPEG-H Audio—The New Standard for Universal Spatial/3D Audio Coding," *J. Audio Eng. Soc.*, vol. 62, no. 12, pp. 821–830 (2014 Dec.). https://doi.org/10.17743/jaes.2014.0049.

[7] G. Theile and H. Wittek, "Principles in Surround Recordings With Height," presented at the *130th Convention of the Audio Engineering Society* (2011 May), convention paper 8403.

[8] H. Lee and C. Gribben, "Effect of Vertical Microphone Layer Spacing for a 3D Microphone Array," *J. Audio Eng. Soc.*, vol. 62, no. 12, pp. 870–884 (2014 Dec.). https://doi.org/10.17743/jaes.2014.0045.

[9] M. Lindberg, "3D Recording With the '2L-Cube,'" http://www.2l.no/artikler/2L-VDT.pdf (accessed Oct. 26, 2021).

[10] K. Hamasaki and W. Van Baelen, "Natural Sound Recording of an Orchestra With Three-Dimensional Sound," presented at the *138th Convention of the Audio Engineering Society* (2015 May), convention paper 9348.

[11] M. Williams, "Microphone Array Design Applied to Complete Hemispherical Sound Reproduction—From Integral 3D to Comfort 3D," presented at the *140th Convention of the Audio Engineering Society* (2016 May), convention paper 9569.

[12] W. Howie and R. King, "Exploratory Microphone Techniques for Three-Dimensional Classical Music Recording," presented at the *138th Convention of the Audio Engineering Society* (2015 May), eBrief 196.

[13] D. Bowles, "A Microphone Array for Recording Music in Surround-Sound With Height Channels," presented at the *139th Convention of the Audio Engineering Society* (2015 Oct.), convention paper 9430.

[14] H. Wittek and G. Theile, "Development and Application of a Stereophonic Multichannel Recording Technique for 3D Audio and VR," presented at the *143rd Convention of the Audio Engineering Society* (2017 Oct.), convention paper 9869.

[15] F. Camerer, "Designing a 9-Channel Location Sound Microphone From Scratch," presented at the *149th Convention of the Audio Engineering Society* (2020 Oct.), eBrief 622.

[16] H. Lee, "Capturing 360° Audio Using an Equal Segment Microphone Array (ESMA)," *J. Audio Eng. Soc.*, vol. 67, no. 1/2, pp. 13–26 (2019 Jan.). https://doi.org/10.17743/jaes.2018.0068.

[17] K. Y.-Y. Zhang and P. Geluso, "The 3DCC Microphone Technique: A Native B-Format Approach to Recording Musical Performance," presented at the *147th Convention of the Audio Engineering Society* (2019 Oct.), convention paper 10295.

[18] H. Lee, "Multichannel 3D Microphone Arrays: A Review," *J. Audio Eng. Soc.*, vol. 69, no. 1/2, pp. 5–26 (2021 Jan.). https://doi.org/10.17743/jaes.2020.0069.

[19] mh acoustics, "Em32 Eigenmike® Microphone Array Release Notes (v17.0)," https://mhacoustics.com/sites/default/files/ReleaseNotes.pdf (accessed Oct. 26, 2021).

[20] Sennheiser Electronic GmbH & Co., "Ambeo VR Mic," https://en-uk.sennheiser.com/microphone-3d-audio-ambeo-vr-mic (accessed Oct. 26, 2021).

[21] RØDE Microphones, "RØDE NT-SF1," https://en.rode.com/ntsf1.

[22] Zylia, "ZYLIA ZM-1 Microphone," https://www.zylia.co/zylia-zm-1-microphone.html.

[23] F. Rumsey, S. Zieliński R. Kassier, and S. Bech, "On the Relative Importance of Spatial and Timbral Fidelities in Judgments of Degraded Multichannel Audio Quality," *J. Acoust. Soc. Am.*, vol. 118, no. 2, pp. 968–976 (2005 Aug). https://doi.org/10.1121/1.1945368.

[24] R. Conetta, T. Brookes, F. Rumsey, et al., "Spatial Audio Quality Perception (Part 2): A Linear Regression Model," *J. Audio Eng. Soc.*, vol. 62, no. 12, pp. 847–860 (2014 Dec.). https://doi.org/10.17743/jaes.2014.0047.

[25] S. George, S. Zielinski, and F. Rumsey, "Feature Extraction for the Prediction of Multichannel Spatial Audio Fidelity," *IEEE Trans. Audio Speech Lang. Process.*, vol. 14, no. 6, pp. 1994–2005 (2006 Nov.). https://doi.org/10.1109/TASL.2006.883248.

[26] H. Lee and D. Johnson, "3D Microphone Array Recording Comparison (3D-MARCo)," (2019 Oct.). https://doi.org/10.5281/zenodo.3477602.

[27] H. Lee and D. Johnson, "An Open-Access Database of 3D Microphone Array Recordings," presented at the *147th Convention of the Audio Engineering Society* (2019 Oct.), eBrief 543.

[28] ITU-R, "Advanced Sound System for Programme Production," *Recommendation BS.2051-2* (2018 Jul.).

[29] G. Theile, "Natural 5.1 Music Recording Based on Psychoacoustic Principals," in *Proceedings of the AES 19th International Conference: Surround Sound - Techniques, Technology, and Perception* (2001 Jun.), conference paper 1904.

[30] H. Wittek, "Image Assistant," https://www.hauptmikrofon.de/stereo-surround/image-assistant (accessed Oct. 26, 2021).

[31] R. Wallis and H. Lee, "The Effect of Interchannel Time Difference on Localization in Vertical Stereophony," *J. Audio Eng. Soc.*, vol. 63, no. 10, pp. 767–776 (2015 Oct.). https://doi.org/10.17743/jaes.2015.0069.

[32] R. Wallis and H. Lee, "The Reduction of Vertical Interchannel Crosstalk: The Analysis of Localisation Thresholds for Natural Sound Sources," *Appl. Sci.*, vol. 7, no. 3, p. 278 (2017 Mar). https://doi.org/10.3390/app7030278.

[33] K. Hamasaki and K. Hiyama, "Reproducing Spatial Impression With Multichannel Audio," in *Proceedings of the AES 24th International Conference: Multichannel Audio, The New Reality* (2003 Jun.), conference paper 19.

[34] A. Heller, R. Lee, and E. Benjamin, "Is My Decoder Ambisonic?" presented at the *125th Convention of the Audio Engineering Society* (2008 Oct.), convention paper 7553.

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

[36] F. Zotter, H. Pomberger, and M. Noisternig, "Energy-Preserving Ambisonic Decoding," *Acta Acust. United Acust.*, vol. 98, no. 1, pp. 37–47 (2012 Jan./Feb.). https://doi.org/10.3813/AAA.918490.

[37] A. Farina, "Advancements in Impulse Response Measurements by Sine Sweeps," presented at the *122nd Convention of the Audio Engineering Society* (2007 May), convention paper 7121.

[38] D. Johnson, A. Harker, and H. Lee, "HAART: A New Impulse Response Toolbox for Spatial Audio Research," presented at the *138th Convention of the Audio Engineering Society* (2015 May), eBrief 190.

[39] mh acoustics, "Auro-3D," https://mhacoustics.com/download (accessed Oct. 26, 2021).

[40] IEM, "IEM Plug-in Suite," https://plugins.iem.at (accessed Oct. 26, 2021).

[41] Aalto University, "Spatial Audio Real-time Applications (SPARTA)," http://research.spa.aalto.fi/projects/sparta_vsts/ (accessed Oct. 26, 2021).

[42] C. Armstrong, L. Thresh, D. Murphy, and G. Kearney, "A Perceptual Evaluation of Individual and Non-Individual HRTFs: A Case Study of the SADIE II Database," *Appl. Sci.*, vol. 8, no. 11, p. 2029 (2018 Oct.). https://doi.org/10.3390/app8112029.

[43] H.-K. Lee and F. Rumsey, "Investigation Into the Effect of Interchannel Crosstalk in Multichannel Microphone Technique," presented at the *118th Convention of the Audio Engineering Society* (2005 May), convention paper 6374.

[44] J. Blauert, "On the Lag of Lateralization Caused by Interaural Time and Intensity Differences," *Audiol.*, vol. 11, no. 5-6, pp. 265–270 (1972 Sep). https://doi.org/10.3109/00206097209072591.

[45] D. W. Grantham and F. L. Wightman, "Detectability of Varying Interaural Temporal Differences[a)]," *J. Acoust. Soc. Am.*, vol. 63, no. 2, pp. 511–523 (1978 Feb). https://doi.org/10.1121/1.381751.

[46] D. Griesinger, "IALF-Binaural Measures of Spatial Impression and Running Reverberance," presented at the *92nd Convention of the Audio Engineering Society* (1992 Mar.), convention paper 3292.

[47] V. Hansen and G. Munch, "Making Recordings for Simulation Tests in the Archimedes Project," *J. Audio Eng. Soc.*, vol. 39, no. 10, pp. 768–774 (1991 Oct.).

[48] P. L. Søndergaard and P. Majdak, "The Auditory Modeling Toolbox," in J. Blauert (Ed.), *The Technology of Binaural Listening*, Modern Acoustics and Signal Processing, pp. 33–56 (Springer, Berlin, Heidelberg, Berlin, Heidelberg, 2013). https://doi.org/10.1007/978-3-642-37762-4_2.

[49] L. R. Bernstein and C. Trahiotis, "The Normalized Correlation: Accounting for Binaural Detection Across Center Frequency," *J. Acoust. Soc. Am.*, vol. 100, no. 6, pp. 3774–3784 (1996). https://doi.org/10.1121/1.417237.

[50] V. Pulkki and M. Karjalainen, *Communication Acoustics: An Introduction to Speech, Audio and Psychoacoustics* (Wiley, Hoboken, NJ, 2015).

[51] D. J. Kistler and F. L. Wightman, "A Model of Head-Related Transfer Functions Based on Principal Components Analysis and Minimum-Phase Reconstruction," *J. Acoust. Soc. Am.*, vol. 91, no. 3, pp. 1637–1647 (1992 Mar.) https://doi.org/10.1121/1.402444.

[52] F. Zotter and M. Frank, "Efficient Phantom Source Widening," *Arch. Acoust.*, vol. 38, no. 1, pp. 27–37 (2013 Mar.). https://doi.org/10.2478/aoa-2013-0004.

[53] D. Griesinger, "Spaciousness and Envelopment in Musical Acoustics," presented at the *101st Convention of the Audio Engineering Society* (1996 Nov.), convention paper 4401.

[54] T. Hidaka, L. L. Beranek, and T. Okano, "Interaural Cross-Correlation, Lateral Fraction, and Low- and High-Frequency Sound Levels as Measures of Acoustical Quality in Concert Halls," *J. Acoust. Soc. Am.*, vol. 98, no. 2, pp. 988–1007 (1995 Aug). https://doi.org/10.1121/1.414451.

[55] P. Zahorik, "Direct-to-Reverberant Energy Ratio Sensitivity," *J. Acoust. Soc. Am.*, vol. 112, no. 5, pp. 2110–2117 (2002 Nov). https://doi.org/10.1121/1.1506692.

[56] A. J. Kolarik, B. C. J. Moore, P. Zahorik, S. Cirstea, and S. Pardhan, "Auditory Distance Perception in Humans: A Review of Cues, Development, Neuronal Bases, and Effects of Sensory Loss," *Attent. Percept. Psych.*, vol. 78,

no. 2, pp. 373–395 (2016 Feb). https://doi.org/10.3758/s13414-015-1015-1.

[57] R. Y. Litovsky, H. S. Colburn, W. A. Yost, and S. J. Guzman, "The Precedence Effect," *J. Acoust. Soc. Am.*, vol. 106, no. 4, pp. 1633–1654 (1999 Oct). https://doi.org/10.1121/1.427914.

[58] F. Zotter and M. Frank, *Ambisonics* (Springer, New York, NY, 2019).

[59] G. Theile, *On the Localisation of Superimposed Soundfield*, Ph.D. thesis, Technische Universität Berlin, Berlin, Germany (1980).

[60] S. P. Lipshitz, "Stereo Microphone Techniques: Are the Purists Wrong?" presented at the *78th Convention of the Audio Engineering Society* (1985 May), convention paper 2261.

[61] C. Gribben and H. Lee, "The Frequency and Loudspeaker-Azimuth Dependencies of Vertical Interchannel Decorrelation on the Vertical Spread of an Auditory Image," *J. Audio Eng. Soc.*, vol. 66, no. 7/8, pp. 537–555 (2018 Jul.).

[62] C. Gribben and H. Lee, "The Perception of Band-Limited Decorrelation Between Vertically Oriented Loudspeakers," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 28, pp. 876–888 (2020 Jan.). https://doi.org/10.1109/TASLP.2020.2969845.

[63] ISO, "Acoustics — Measurement of Room Acoustic Parameters — Part 1: Performance Spaces," *ISO 3382-1:2009* (2009).

[64] E. Larsen, N. Iyer, C. R. Lansing, and A. S. Feng, "On the Minimum Audible Difference in Direct-to-Reverberant Energy Ratio," *J. Acoust. Soc. Am.*, vol. 124, no. 1, pp. 450–461 (2008 Jul.). https://doi.org/10.1121/1.2936368.

THE AUTHORS

Hyunkook Lee

Dale Johnson

Hyunkook Lee is a Reader (i.e., Associate Professor) in Music Technology and the Director of the Applied Psychoacoustics Laboratory (APL), University of Huddersfield, UK. His recent research has advanced understanding about the psychoacoustics of vertical stereophonic localization and spatial impression in 3D sound recording and reproduction. This provided theoretical bases for the developments of several 3D microphone arrays, including Schoeps ORTF-3D. His current research topics include the six-degrees-of-freedom perception and rendering of virtual acoustics and the measurement of multimodal immersive experience for extended reality applications. From 2006 to 2010, he was a Senior Research Engineer in audio R&D with LG Electronics in South Korea, where he participated in the standardizations of MPEG audio codecs and developed spatial audio algorithms for mobile devices. He received a bachelor's degree in music and sound recording (Tonmeister) from the University of Surrey, UK, in 2002 and a Ph.D. in spatial audio psychoacoustics from the Institute of Sound Recording at the same University in 2006. He is a Fellow of the AES and the Vice Chair of the AES High Resolution Audio Technical Committee.

•

Dale Johnson is a research fellow in the Applied Psychoacoustics Laboratory (APL) at the University of Huddersfield. From 2014 to 2018, he studied for a Ph.D. after graduating with a First Class degree in BSc Music Technology and Audio Systems (Hons). His Ph.D. focused on the development of a perceptually motivated optimization algorithm for artificial reverb. During his time with the APL, since 2018 he was involved in the Innovate UK-funded Volumetric Audio Synthesis for AR project (VASAR), in which a real-time virtual acoustics rendering system for six-degrees-of-freedom AR audio experiences was developed. His background is in music, audio engineering, and software development, and his current research focuses on the recording and reproduction of 3D audio for VR and AR and the development of artificial reverberation algorithms.