3D Microphone Array Recording Comparison (3D-MARCo): Objective Measurements

Hyunkook Lee and Dale Johnson

Applied Psychoacoustics Lab (APL), University of Huddersfield, Huddersfield, HD1 3DH, United Kingdom

Correspondence should be addressed to Hyunkook Lee (h.lee@hud.ac.uk)

ABSTRACTS

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 0m and 1m 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.

Keywords: 3D microphone array; 3D sound recording; objective measurement

0 INTRODUCTION

Three-dimensional (3D) audio is rapidly becoming a new standard for audio content production, delivery and reproduction. New formats such as Dolby Atmos [1], Auro-3D [2], DTS:X [3], NHK 22.2 [4] and Sony 360 Reality Audio [5], along with the recently standardised MPEG-H [6] 3D audio codec, 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 [7-18] – a comprehensive review of existing 3D microphone arrays is provided in [19]. Furthermore, with the burgeoning interest in head-tracked binaural audio for extended reality applications, Ambisonic microphone systems [e.g., 20-23] are used more widely than in the past for its convenience in sound field rotation.

With an increasing number of available 3D acoustic music recordings, there arises the need for evaluating the qualities of such recordings in a systematic way. Much research has been undertaken on the quality evaluation of horizontal-only surround sound recording (e.g., [24-27]). However, research on the perceived qualities of surround recordings with the so-called 'height' channels is still in its early stage. Several studies have compared different 3D microphone techniques (see Sec. 2 in [19] for a review). They generally suggest that different techniques had different pros and cons depending on the tested attributes. However, as pointed out in [19], they 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 systematic and comprehensive investigation into the perceptual characteristics of different 3D microphone arrays, it would be first 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 minimise the influence of recording systems, which would allow for a more controlled comparison on microphone-array-dependent spatial and timbral qualities. Such a database of 3D recordings has recently been created by the present authors [28,29] and named '3D-MARCo' (3D Microphone Array Recording Comparison). 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 9-channel or 8-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 thirteen different source positions using all of the microphones to allow for objective analyses of the microphone arrays as well as the creation of virtual sound sources for future experiments.

As the first step towards to a series of planned formal evaluations of the 3D microphone arrays included in the database, the present study measures various objective parameters in order to gain insights into physical differences among the microphone arrays. The results from this investigation are expected to serve as bases for hypothesising and explaining perceptual differences among the arrays, which will be investigated formally in future subjective listening tests. The rest of the paper is organised as follows. Section 1 briefly summarises recording techniques for the 3D-MARCo database. Section 2 describes the objective parameters and the methods used for computing them. Section 3 then presents and discusses the results.

1 MICROPHONE ARRAYS AND RECORDING SETUP

A total of seven different microphone arrays from the 3D-MARCo database [28,29] were compared in the present study. Stimuli for the objective measurements conducted were created using multichannel impulse responses captured using the microphone arrays. This section briefly describes the array configurations as well as the method used for the impulse response acquisition. Full details about the database are available in [28,29].

This paper uses the following channel labelling and loudspeaker angles for reproduction in Table 1. ITU-R BS.2051-2 [30] recommends the labelling of channels according to the layer and the loudspeaker azimuth angle (e.g., M+060, U-135, B+045, etc.). This would be more useful for referring to various systems including a large number of channels in different layers including the

bottom layer. However, the current investigation deals with only nine channels for reproduction, and therefore it was decided to use the simpler labels as in Table 1. The azimuth and elevation angles of the loudspeakers were chosen based on ITU-R BS.2051-2 [30]. This configuration is also in line with typical loudspeaker layouts for 9-channel 3D home-cinema systems, such as Dolby Atmos 5.1.4 and Auro-3D 9.1.

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

Channels	Labels	Azi. (deg)	Ele. (deg)
Front Left	FL	+30	0
Front Right	FR	–30	0
Front Centre	FC	0	0
Rear Left	RL	+120	0
Rear Right	RR	–120	0
Front Left height	FLh	+45	+45
Front Right height	FRh	–45	+45
Rear Left height	RLh	+135	+45
Rear Right height	RRh	–135	+45

1.2 Microphone Arrays

Table 2 lists and categorises 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 Appendix.

Table 2. 3D microphone arrays included in the 3D-MARCo database, classified according to [19].

	Perceptually motivated		**Physically motivated**
	Horizontally and Vertically Spaced (HVS)	**Horizontally spaced/ vertically coincident (HSVC)**	**Horizontally and Vertically Coincident (HVC)**
Main array	OCT-3D 2L-Cube Decca Cuboid	PCMA-3D	Eigenmike EM32
Ambience array	Hamasaki Square (HS) with height layer at 1m above	Hamasaki Square (HS) with height layer at 0m	

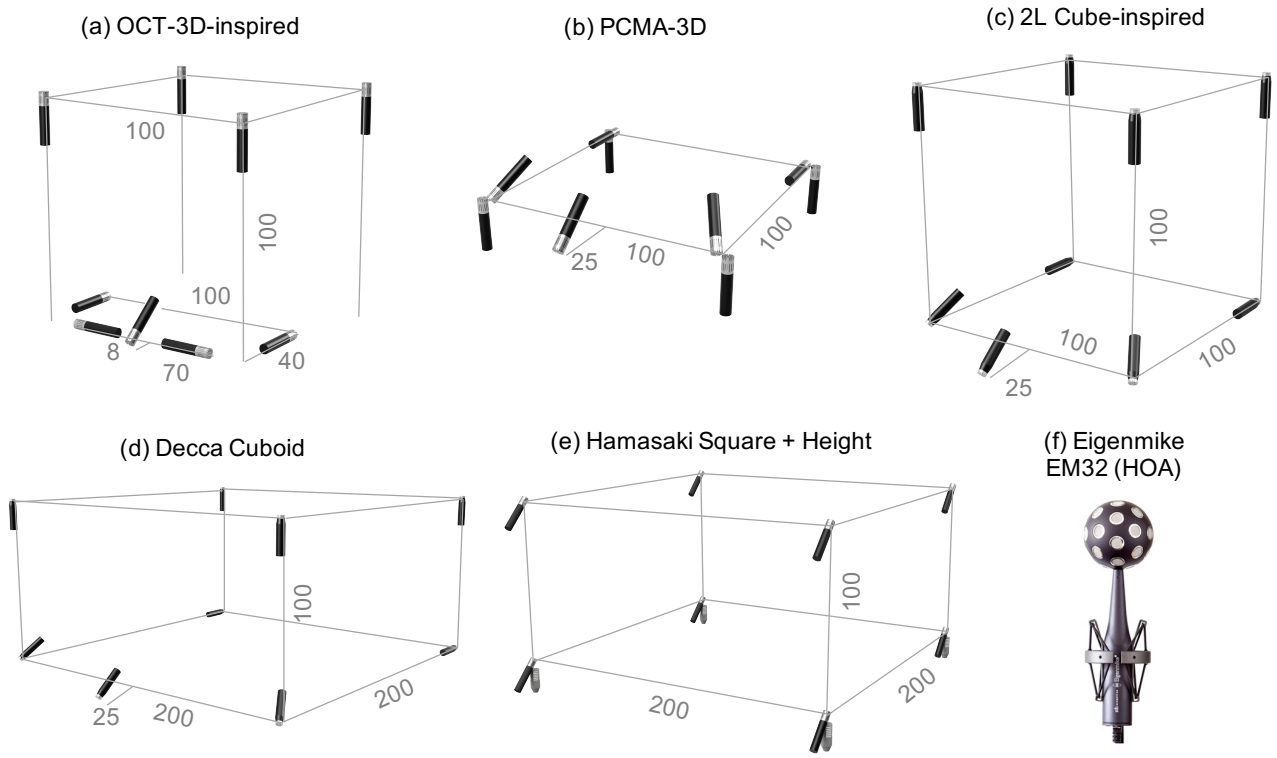

Fig. 1. Microphone arrays used for the recording and objective measurements. Unit for the numbers is cm. All microphones except for the Eigenmike EM32 and the Hamasaki Square (Schoeps CCM8) were of the DPA d:dicate series. Detailed information on the polar patterns and microphone angles for each array can be found in Appendix.

1.2.1. OCT-3D

OCT-3D (Fig. 1a), proposed by Theile and Wittek [7], augments the OCT (Optimised Cardioid Triangle)-surround 5-channel microphone array [31] with four upward-facing supercardioid microphones placed 1m above the base layer. The main design goal of OCT is to minimise interchannel crosstalk (ICXT) for accurate frontal image localisation [31]. The front triplet uses a cardioid centre microphone placed 8cm in front 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 70cm spacing was used to produce the SRA of 115° [32]. The rear microphones were backward-facing cardioid microphones with 1m spacing, placed at 40cm 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 from the front centre one. However, in the 3D-MARCo session, the height layer was modified to be of a 1m x 1m square to be consistent with the PCMA-3D's height layer.

1.2.2 PCMA-3D

The PCMA-3D (Fig. 1b) is based on the 'Perspective Control Microphone Array' design concept [33], which allows a flexible rendering of perceived distance in 5-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 [34], (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 [35]. 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 upwards in order to suppress the ICXT maximally.

1.2.3 2L-Cube

2L-Cube is a technique developed by Lindberg [9]. It employs nine omni-directional microphones in a 1m x 1m x 1m cube arrangement, thus mainly relying on interchannel 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 2LCube 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 1m x 1m height layer at 1m vertical spacing.

1.2.4 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 = 2m to 2.5m, FC–base = 1m to 1.5m), thus heavily relying on ICTD for phantom imaging. In 3D-MARCo, the traditional Decca Tree was augmented with rear microphones placed at 2m behind the base point and height microphones 1m above the base layer, thus named 'Decca Cuboid' here. The horizontal dimensions of this array are twice as large as 2L-Cube whilst keeping the vertical dimension the same. Therefore, a greater amount of interchannel decorrelation can be expected. The FC microphone was placed 0.25m in front of the base point instead of the originally used 1m. 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 to avoid too strong a centre image.

1.2.5 Hamasaki Square with Height

Hamasaki Square [6] is a popular technique for recording 4-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 2m x 2m square. The height layer employed four cardioid microphones at two vertical positions from the base layer for a comparison purpose: 0m (i.e., vertically coincident based on [8]) and 1m (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 0m height layer.

1.2.6 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 a different order between 1 and 4 for Ambisonic reproduction. In the current study, the 1st and 4th order Ambisonic reproductions were compared. Although an ideal Ambisonic reproduction requires a loudspeaker

array configured in a regular polygon or polyhedral layout [37], 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 [30]), using decoders optimised for the purpose (e.g., ALLRAD [38] and EPAD [39]). Although the main focus of the current investigation is the reproduction of perceptually motivated microphones that were developed for an ITU-R-based 9-channel loudspeaker playback, Eigenmike EM32 was also included for a comparison purpose as in practice Ambisonic recordings might be reproduced over such an irregular loudspeaker array more frequently than an ideal regular array.

1.3 Multichannel Room Impulse Responses

Multichannel room impulse responses (MRIRs) were captured for thirteen source positions (Fig. 2) using all of the microphone arrays described above. The recording venue was St.Paul's concert hall in Huddersfield, UK, which is a converted church with the dimensions of 16m (W) x 30m (L) x 13m (H). The average RT60 of the hall is 2.1s.

The exponential sine sweep method [40] offered by the HAART software [41] was used for the acquisition of the MRIRs. Genelec 8331A co-axial loudspeakers were used as sound sources. Their acoustic centre was at 1.14m above the floor. The loudspeakers were positioned with 15° intervals from -90° to 90°. The distance from the base point of the mic arrays to each loudspeaker was 3m for +90°, +60°, +30°, 0°, –30°, –60° and –90°, and 4m for +75°, +45°, +15°, –15°, –45° and –75°.

The objective measurements described in the following sections produce an extensive amount of data to be analysed for each source position. Hence, for the present study, only the intermediate source position +45° were used (see Fig. 4). 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 localisation and spatial impression.

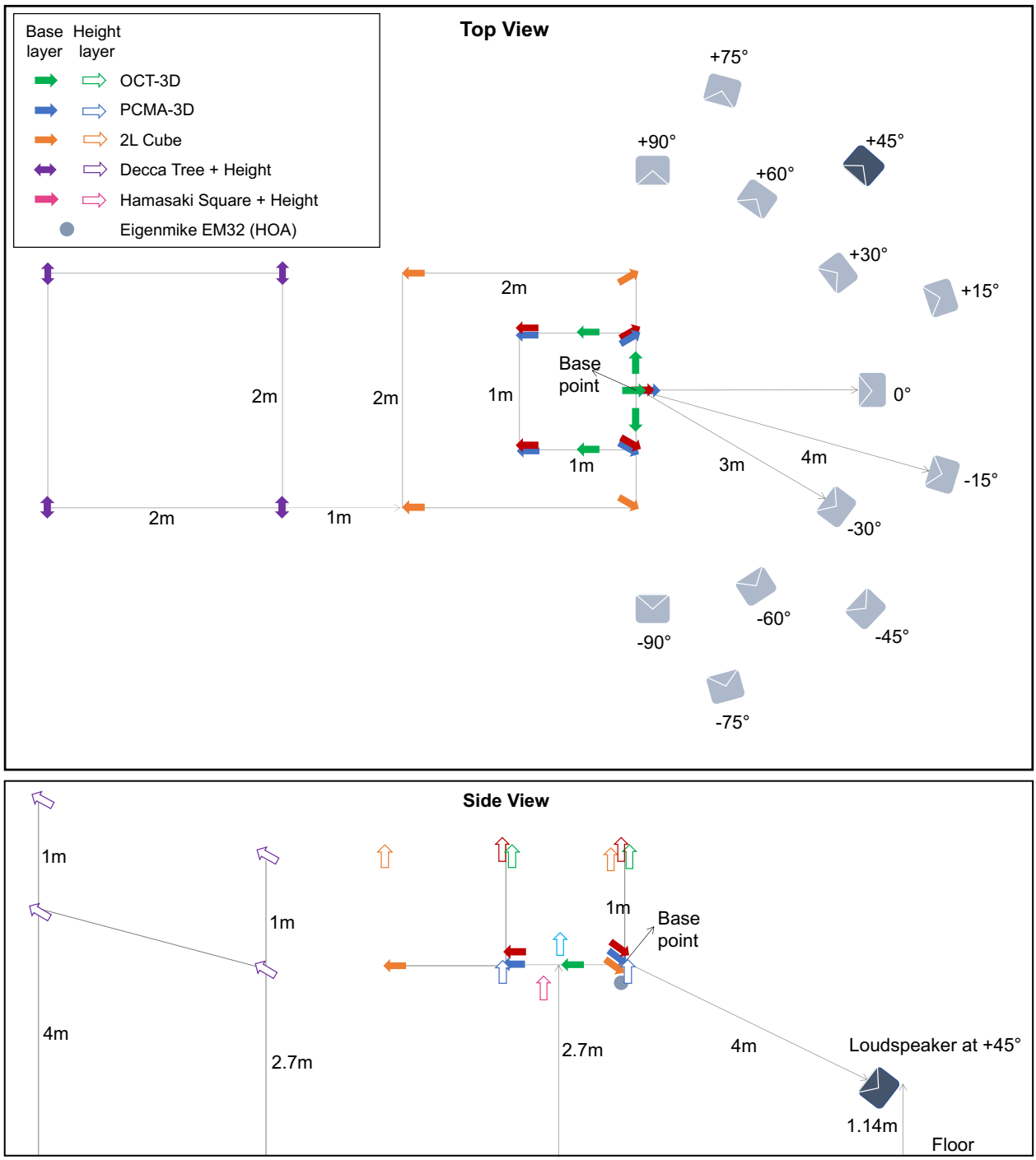

Fig. 2. Physical layout of the microphones and loudspeakers used for capturing the multichannel room impulse responses (MRIRs) in 3D-MARCo. For the present objective measurements, the MRIRs for the source at +45° were used.

2 OBJECTIVE MEASUREMENTS

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

- Interchannel level difference (ICLD) and interchannel time difference (ICTD) of interchannel crosstalk (ICXT).
- Fluctuations in 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).

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 colouration, 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 synthesised 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 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. 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 [42], which were then decoded to the loudspeakers configured as in Table 1. The ALLRADecoder plugin in the IEM plugin suite [43] was used since the ALLRAD method [38] used in the plugin is specifically designed for decoding an Ambisonic recording to irregular loudspeaker arrays such as the one used here (i.e., Table 1). The decoder weighting option in the plugin was set to 'basic', which is optimised for an ITD synthesis in reproduction at frequencies below around 700 Hz [37]. 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 optimised 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 Ambisonic decoders. Readers who are interested in exploring various decoding options are recommended to use the IEM [43] or SPARTA [44] plugin suite on the Reaper session template provided with the 3D-MARCo database [29].

The loudspeaker signals were either kept as broadband or split into different frequency bands, depending on the parameters measured. The BIRRs were synthesised by convolving the MAIRs with the KU100 head-related impulse responses (HRIRs) taken from the SADIE II database [45]. The MAIRs or BIRRs underwent time-window segmentation as required for each of the parameters. Detailed descriptions for the segmentation are provided in each subsection below.

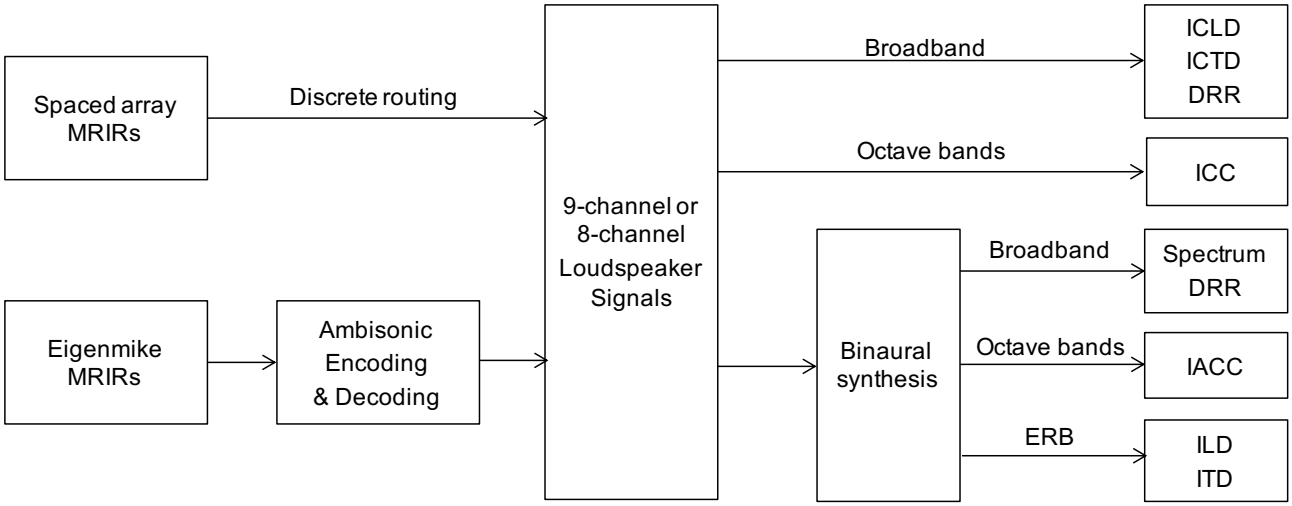

Fig. 3. Overall workflow for the objective measurements conducted.

2.2 Objective Measures

2.2.1 Level and Time Differences to 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 localisation of phantom image. Research suggests that horizontal ICXT is significantly associated with perceptual effects such as locatedness (i.e., ease of localisation) and source image spread [46]; for the frontal three microphones in the base layer, a high level of ICXT tends to decrease locatedness and increase HIS, and the magnitude of this effect becomes greater with a larger time delay of ICXT. Between vertically oriented microphones, on the other hand, ICXT present in the height microphone signal (e.g., FLh) would cause the phantom source to be shifted upwards regardless of ICTD if it is not suppressed by at least 7 dB in reference to the direct sound in the base microphone signal (e.g., FL) [35].

In the current measurement, FL was taken as a reference channel that an ICTD was calculated against 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 ICTD of each signal to FL was calculated as the lag (in ms) of the maximum absolute value of the normalised cross-correlation function (NCF) (Eq. 1), using the MRIRs.

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

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 = 10ms). The lag (τ) 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 coherent signals has a potential risk of introducing a greater degree of spectral distortion in the ear-input signal due to the comb-filter effect. The height microphone layer in concert hall recordings primarily aims to provide extra ambience to enhance spatial impression, whereas the base layer focuses on sound source imaging. However, not only the base layer but also the height layer picks up a certain level of direct sounds (i.e., ICXT) with different ICTDs, depending on their polar patterns and configuration. Therefore, when all of the signals are summed at the ear, the ICXT in the height layer signals might affect the frequency responses of the ear-input signals of the main layer, thus potentially influencing the perceived tonal characteristics of source images.

To investigate the spectral influence of the height layer 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 amongst all arrays was 9.5 ms for RRh–FL of Decca Cuboid (Fig. 4(b)).

2.2.3 Fluctuations in 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 been found to be related to the perceived movement or spread the image depending on the fluctuation rate (i.e., the “localisation lag” phenomenon [47]). That is, at low rates of fluctuations (up to 3–20 Hz, depending on the experimental method and the type of signal [47,48,49]), 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 a 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 [50] 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 a 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 [51]. Half-wave rectification and a 1st-order low-pass filtering at 1 kHz were applied to mimic the breakdown of the phase-locking mechanism as used in [52,53]. The resulting signals were then time-segmented into 50%-overlapping Hann-windowed 50ms 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 ±1ms [54]. 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 centre frequency of 1.47 kHz and below were averaged, whereas the ILDs were averaged for the ERBs with the centre 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 reproduction as well as listener envelopment (LEV) [36,55,56]. It is also related to the size of listening area (i.e., the more decorrelated, the wider the 'sweet spot') [36]. For the present investigation, interchannel correlation coefficient (ICC) is calculated as the absolute value of the NCF (Eq. 1) with zero lag. 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. For computing ICCs, the MRIRs were first split into nine octave bands with their centre frequencies ranging from 63 Hz to 16 kHz, using an 8^{th}-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 = 2100 ms) in order to predict differences in source-related and environment-related spatial attributes [57]. The 80ms boundary point between the two segments is typically used for musical sources in concert hall research [58]. 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)$_{High}$. As with ICXT, ICC was computed for each of the microphone signals against FL. Additionally, the symmetrically arranged microphones pairs RL-RR, FLh-FRh and RLh-RRh were included in the measurements.

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. Hidaka et al. [58] found that ASW and LEV in concert halls were best predicted using the average of the IACCs for the octave bands with the centre 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), whilst 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 synthesised 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

The direct-to-reverberant energy ratio (DRR) is widely known as an absolute measure for perceived auditory distance in rooms [59]. 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 is short enough to exclude the first reflection [61]. 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 to 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 [31], 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 and based on [38], 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 [61] in combination with the ICLD, thus locating the phantom image in the front, 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 localisation.

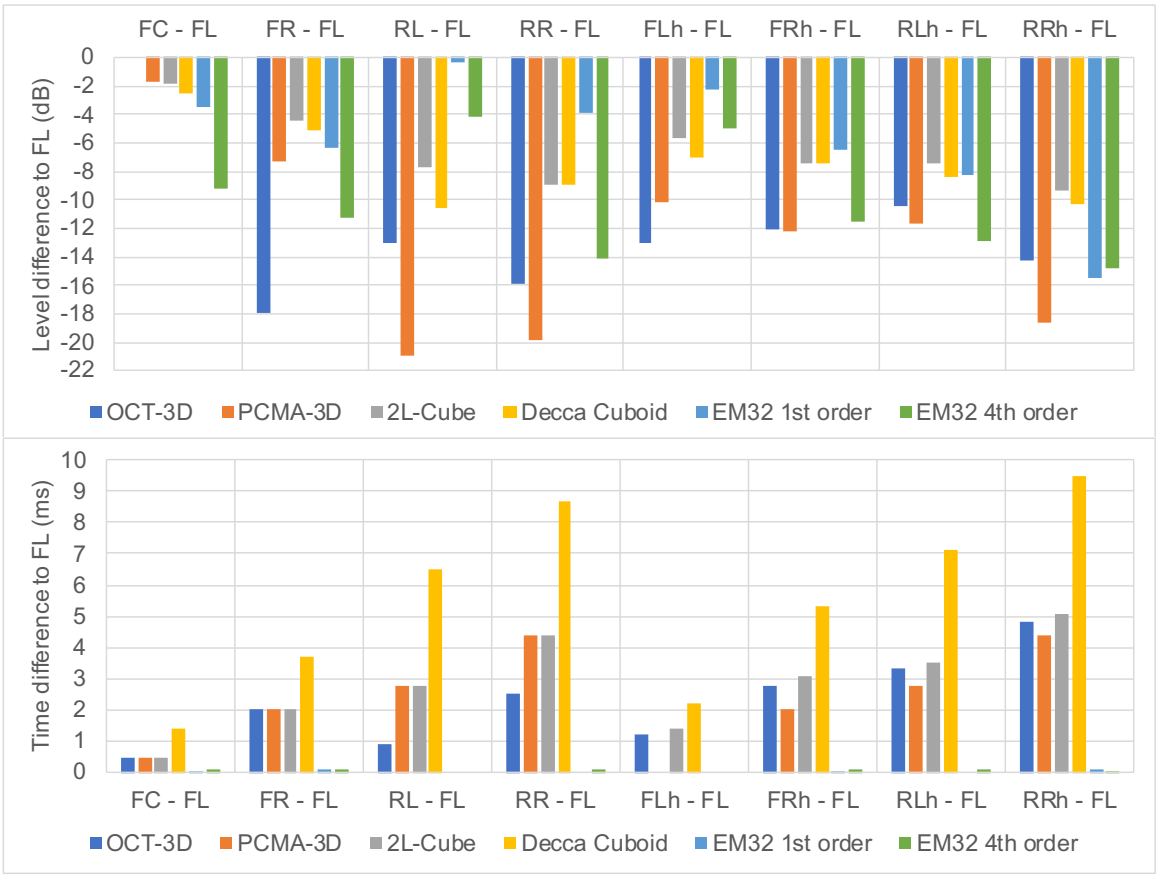

Fig. 4. Interchannel level and time differences (ICLD and ICTD) of each microphone to FL, measured using the energy of the direct sound portion (0–2.5 ms) of the impulse responses captured for the +45° source position. ICLDs were not calculated for the Hamasaki Square arrays as their aim is to capture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

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 localisation between the horizontal channels. However, as reported in [34], the precedence effect would not operate between vertically oriented loudspeakers by ICTD alone. That is, when the levels of the lower and upper loudspeakers are the same, the phantom image would not be localised at the position of the lower loudspeaker even if the upper loudspeaker is delayed more than 1ms, but perceived at a random position depending on the spectrum of the ear-input signal. As mentioned earlier, at least a reduction of 7 dB would be required to avoid the localisation uncertainty. 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 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 localisation 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 upwards. Note that the omni height microphones of 2L-Cube and Decca Cubioid were facing directly upwards in the recording session. If the microphones had been facing directly towards the sound source, then the level of ICXT would have been higher, which might increase the strength of its perceived artifacts.

The Eigenmike conditions generally show that the 4th order rendering had a considerably lower level of ICXT than the 1st order rendering, which was an expected result due to the increased spatial resolution of the higher-order Ambisonics. The channel separation of the 1st order was found to be particularly small for RL-FL (–0.3 dB) and FLh-FL (–2.3 dB). In contrast with the other arrays that are perceptually motivated, 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 [62].

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 upwards-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. Although PCMA-3D also has ICTDs between diagonally oriented main and height microphones, e.g., FL-FRh, the resulting comb-filter effect would be weak owing to the low level of ICXT in the height signals. The comb-filter pattern observed at frequencies above 8 kHz in the delta plot for PCMA-3D seem to be due to the slight gap between the diaphragms that existed inevitably due to the microphone enclosure.

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 colouration 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.

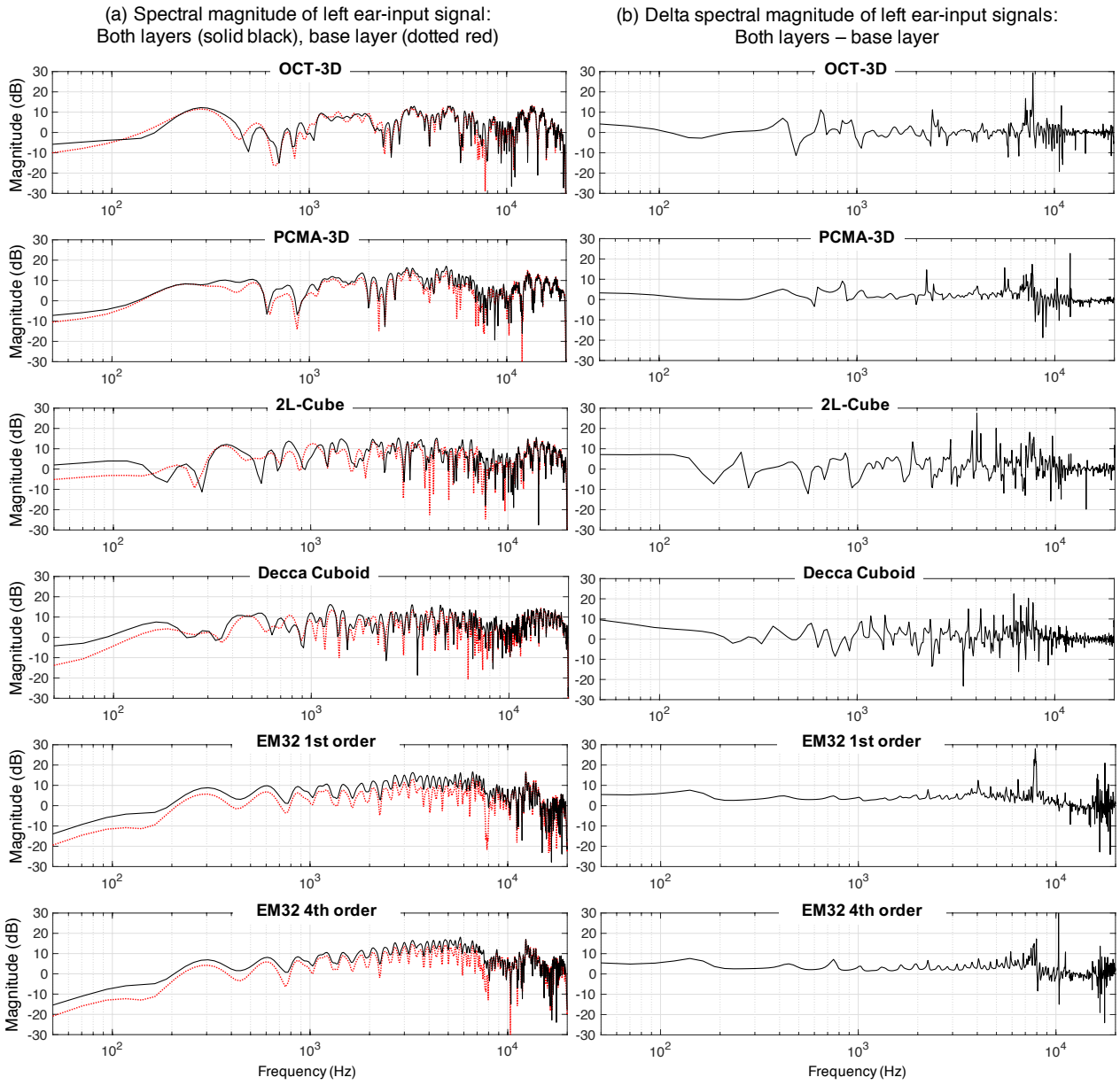

Fig. 5. Spectral magnitudes of the left-ear input signal of the binaural impulse responses resulting from the loudspeaker playback of multichannel impulse responses. (a) measurements for both the base and height layers and those for the base layer only, (b) difference of both layers to the base-layer-only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

Observing the magnitude spectra of both layers in Fig. 5(a), 2L-Cube and Decca Cuboid had more energy below about 150 Hz than the other arrays. This was expected since omni-directional microphones tend to have a more extended frequency response than uni-directional ones. However, 2L-Cube and Decca Cuboid also appear to have considerably less energy between 200 Hz to 400 Hz, and generally more complex spectrum compared to the others. The Eigenmike conditions had the least amount of low frequency energy amongst all of the arrays. However, their overall frequency responses were most even owing to the coincident nature, despite the comb-filter pattern due to the floor reflection at 2.8 ms that was included in the time window.

The above results imply potentially substantial differences among the arrays in perceived tonal colour. However, the subjective interpretation of tonal colour seems to be a complex cognitive process, which may depend on the type of sound source but also be related to one's experience and expectation. For example, in a standard 2-channel reproduction with loudspeakers, comb-filtering is always present in the ear signals due to the interaural crosstalk. However, we do not necessarily perceive such spectral distortion as tonal colouration hypothetically because the brain might be highly familiar with the pattern. Similarly, tonal colour perception in 3D reproduction may also be related to what the listener is familiar with in terms of the types of sound source and production method. Furthermore, Theile's 'association model' [63] suggests that the perception of the tonal colour of a phantom image is also related to localisation; the audibility of tone colouration 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. The black and red plots show the results for the pink noise and trumpet sources. To quantify the magnitude of fluctuation, three standard deviations (3SD) are presented 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 [64]'s observation on 2-channel stereo microphone techniques. Furthermore, since 2L-Cube and Decca Cuboid had a great amount of ICXT than PCMA-3D and OCT-3D, it can be also suggested that an array with a greater amount of ICXT would cause a greater magnitude of ITD fluctuation.

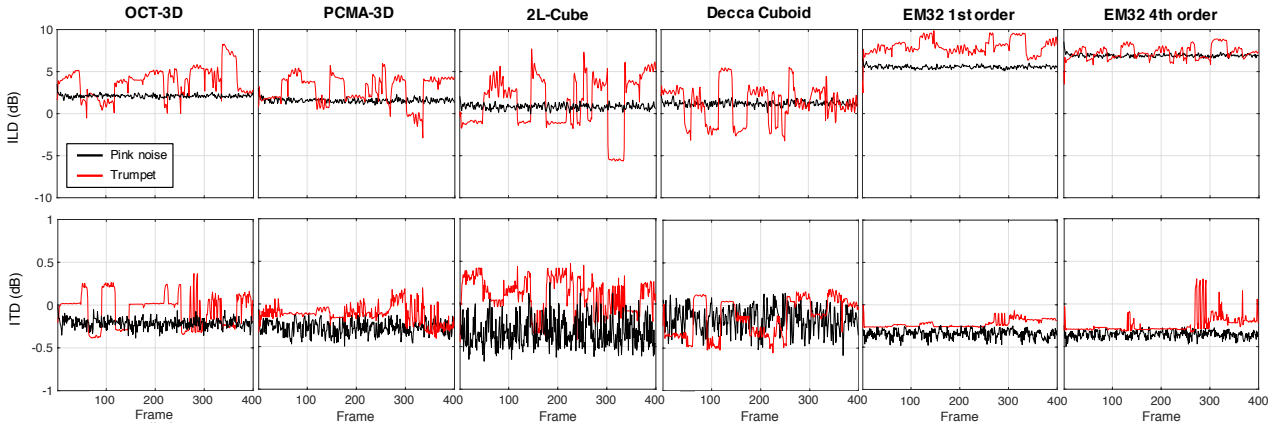

Fig. 6. ILDs and ITDs measured for the 50%-overlapping 50ms Hann-windowed frames of 10-second-long pink noise (black) and anechoic trumpet (red). The ILD and ITD for each frame are the averages of ILDs and ITDs computed for the ERBs with the centre frequencies between 1.62 kHz and 19 kHz and for those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

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 the studies using pulse train or modulated noise [47,48,49]. 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 [47,48,49].

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 OCT-3D, 2L-Cube and Decca Cuboid, which are in the HVS category, the ILDs and ITDs had large occasional shifts between positive and negative values.

PCMA-3D, which is a 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 amongst 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, thus a poorer imaging stability.

Table 3. means and three standard deviations (3SDs)

Array	Noise				Trumpet			
	ILD (dB)		ITD (ms)		ILD (dB)		ITD (ms)	
	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

3.4 Interchannel Cross-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 only exception was the vertically coincident FL-FLh condition for PCMA-3D that produced considerably high ICCs for all bands. 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. This again seems to be due to the coincident nature of the microphone system.

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 seems reasonable considering the larger microphone spacing of 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 ±90°-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 whilst suppressing those from the left-hand side, which would eventually have lowered the ICC. Conversely, the omni-directional FL and FR of 2L-Cube would have captured early reflections from both sides with little level difference. PCMA-3D uses cardioids for FL and FR, but their subtended angle from the centre line was 30°, which would not be large enough to separate the early reflections captured by the microphones to a large degree. On the other hand, the differences among the three arrays in ICC L_{Low} were much smaller than those in ICC E_{Low}, perhaps due to the random nature of diffuse reverberation.

It is interesting to observe that the front-rear microphone pairs FL-RL and FL-RR had an opposite pattern to the FL-FR discussed above. That is, both ICC E_{Low} and ICC L_{Low}, OCT-3D was the most correlated among the spaced arrays, with PCMA-3D (0.17) being more slightly decorrelated than 2L-Cube, which had the same horizontal array size. This seems to be because PCMA-3D not only had a weaker ICXT, but also had a larger ICTD than OCT-3D in RL and RR. PCMA-3D also had a weaker ICXT in RL and RR than 2L-Cube, whereas their ICTDs were the same. Decorrelation between the front and rear channel signals in surround reproduction may be considered to be associated with perceived lateral image spread or auditory depth, which requires further research. Despite the differences discussed above, the ICCs of all of the horizontally spaced arrays for FL-RL and FL-RR seem to be low enough to avoid any unpleasant phasiness during head movement.

Fig. 7. Interchannel correlation coefficients for different pairs of microphone signals.

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 1m spacing between the microphones and therefore their ICCs were comparable to those of the other spaced main arrays in general. Gribben and Lee [65] found that in a 9-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 hypothesised that, if any differences in perceived VIS were perceived among the spaced main arrays, it would be due to ICXT rather than ICC.

Griesinger [66] 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 and Eigenmike 1st order had the lowest (0.19) and highest (0.63) values amongst all, respectively. The difference between PCMA-3D (0.36) and 2L-Cube (0.34) was negligible, whilst OCT-3D had a slightly higher ICC L (0.44) than them. A similar pattern was found for RLh-RRh, except that OCT-3D, PCMA-3D and 2L-Cube did not have any meaningful difference and Eigenmike 4th order had the highest value. From these results, it could be predicted that the perceived magnitude of LEV would be correlated with the horizontal microphone spacing.

For the Eigenmike conditions, it appears that the difference between the 1st and 4th orders generally became larger with an increasing frequency band, depending on the channel pair. For instance, the 4th order had a dramatic decrease of ICC E from 0.67 to 0.1 for FL-FR as the band increased from low to high, whilst the 1st 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 1st and 4th orders regardless of the frequency band. This might suggest that, in the current 9-channel loudspeaker reproduction, the well-known limitation of Ambisonic loudspeaker reproduction regarding phasiness during head movement would still exist even at the higher order.

However, 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 optimised for an ITD synthesis in reproduction at frequencies below around 700 Hz [37]. The result might be different if the decoder used the 'max rE' weighting, which is optimised 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.

The ambience arrays HS-0m and HS-1m generally had lower ICCs than the main arrays at the low bands, whereas the differences were negligible for most channel pairs in both segments. However, the ICCs for the main and ambience arrays for the early segment might have different perceptual effects. Since the DRRs of all of the HS array signals were much lower than those of the main array signals, the ICC Es of the HS arrays would be mainly determined by early reflections, whereas those for the main arrays would be influenced by ICLD and ICTD of the direct sound. Therefore, the ICCs for the main arrays would be associated with source-related attributes such as ASW, perceived source distance and loudness, whereas those for HS would affect the perception of more environment-related width and depth attributes.

3.6 Interaural cross-correlation coefficient (IACC)

The results are plotted in Fig. 8 (a) to (c). Additionally, 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. In general, the IACC E3 values for the Eigenmike conditions were higher than those for the horizontally spaced arrays for all layer conditions, following a trend similar to the ICC results. However, their differences in the results for IACC L3 appear to be smaller. The 4th order Ambisonic condition even had a slightly smaller IACC L3 than some of the spaced arrays. This result seems to suggest that the differences between the spaced and coincident arrays would be larger in ASW rather than in LEV.

It can be also observed that differences among the spaced main arrays (OCT-3D, PCMA-3D, 2L-Cube and Decca Cuboid) 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,

suggesting smaller differences in ASW. This is mainly due to the decrease in IACC E3 for OCT-3D (–0.15) and the increase for 2L-Cube (0.1) and Decca Cuboid (0.05) when the height layer was added. Although these changes are only small, their effect on ASW may still be slightly audible since the just noticeable difference (JND) of ASW is known to be 0.075 [57]. PCMA-3D was hardly influenced by the height layer in IACC E3.

Although IACC L3 for the height-layer-only condition was considerably higher than that for the base-layer-only in general, when the both layers were present, 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. This suggests that LEV might be determined mainly by the correlation between the ear signals resulting from the base layer rather than that from the height layer.

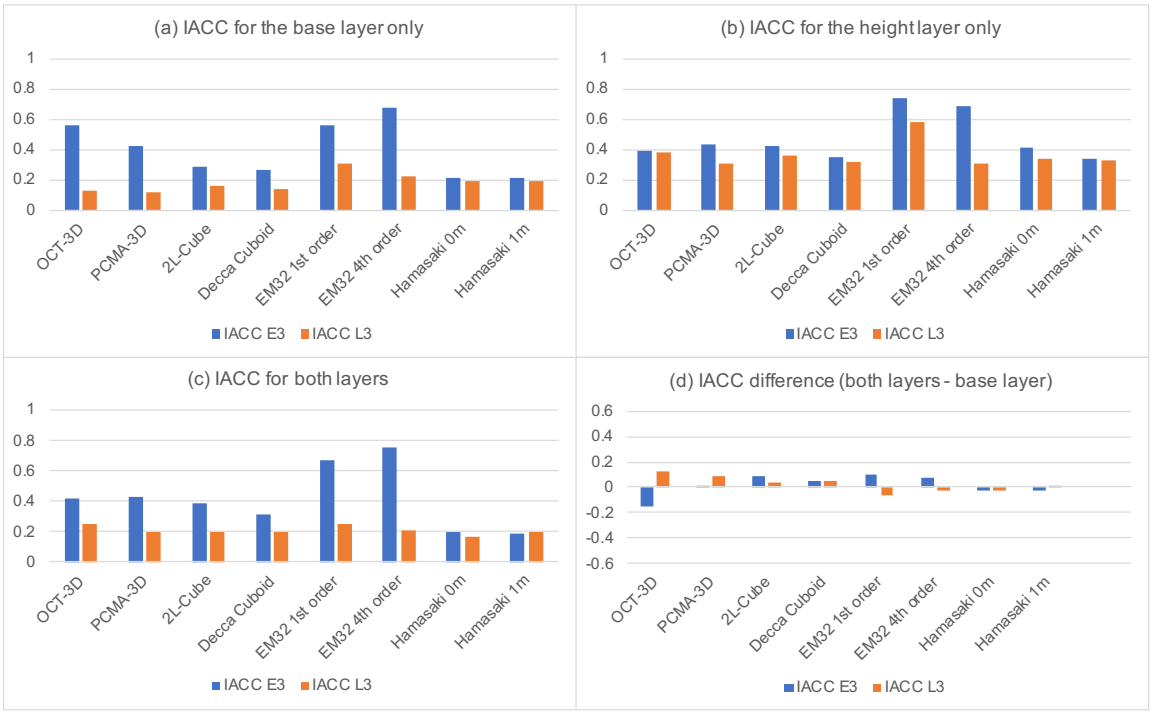

Fig. 8. Interaural cross-correlation coefficients (IACCs) for ear-input signals resulting from different microphone signals reproduced from a binaurally synthesised 9-channel 3D loudspeaker system.

Another interesting result that can be observed is that the two vertical spacings of 0m and 1m 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.

3.7 Direct-to-Reverberant Energy Ratio (DRR)

Fig. 9 shows the measurement results. At a glance, it is obvious that the Hamasaki Square signals had the lowest DRRs in general. The negative values indicate that the direct sound energy was smaller than the reverberant energy as intended for the ambience array. For individual channel signals, differences between the different arrays 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 (–8dB) 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.

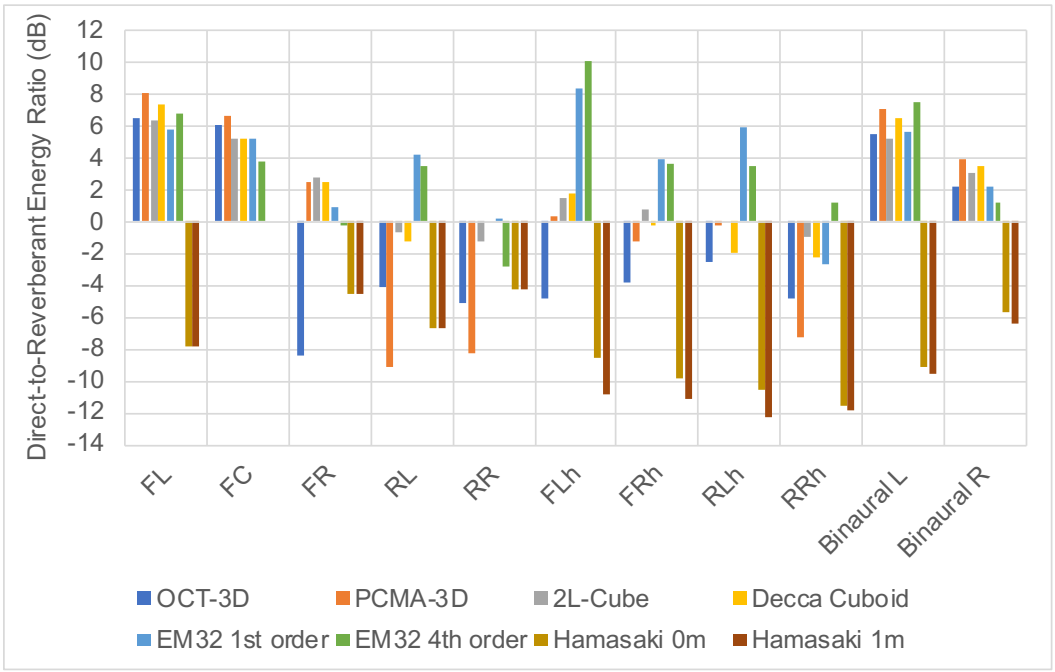

Fig. 9. Direct-to-Reverberant Ratio (DRR) for each microphone and ear-input signal.

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 2.4 dB between 2L-Cube and Eigenmike 4th for the left ear, and 2.7 dB between Eigenmike 4th and PCMA-3D for the right ear. The difference between HS 0m and HS 1m was only 0.3 dB and 0.7 dB for the left and right ears, respectively. 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. [67] 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 [60] 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. However, it is not clear yet whether it is the DRR of the binaural signals or a channel-dependent weighting of DRR that affects 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 0m and 1m 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, the spectral influence of interchannel crosstalk, fluctuations of interaural level and time

differences, interchannel cross-correlation coefficient, interaural cross-correlation coefficient, and direct-to-reverberant energy ratio. The aim of these measurements was to produce theoretical 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 interchannel crosstalk, and this was found to be related to the considerable differences in the amount of spectral distortion in the ear signal as well as in the magnitude of ILD and ITD fluctuation over time. From this, it is expected that the arrays would have audible differences in perceived timbral characteristics as well as the localisation stability and spread of phantom image.

The arrays would have considerable differences in the perceived magnitudes of horizontal spatial impression and the size of listening area due to the large differences in interchannel decorrelation between horizontal channels. Considerable differences in vertical interchannel decorrelation were also observed. However, based on previous research findings, this is hypothesised to have a minimal effect on perceived vertical image spread.

The analysis of interaural cross-correlation suggests that the addition of the height layer to the base layer would have a minor effect on ASW and LEV regardless of the array type, even though the base and height layers might have audible differences independently.

The differences between the microphone arrays in the D/R ratios of ear-input signals resulting from the virtual 9-channel loudspeaker reproduction were around or below the just noticeable difference of perceived auditory distance (i.e., 4 dB), even though the D/R ratios of individual microphone signals had larger differences among the arrays. This raises an interesting question as to whether it would be the channel-dependent D/R ratio or the D/R ratio of the final ear signal that affects perceived auditory distance.

Further works will include the elicitation of perceptual differences among the microphone arrays in order to establish a set of defined attribute scales, which will then be used for a grading experiment. This will allow for analysing the relationship between the subjective results and the objective results from the present paper. From this, the perceptual weightings of the objective

parameters on the subjective ratings will be determined to develop a prediction model for 3D acoustic recording quality evaluation.

5 ACKNOWLEDGMENT

This project was partly funded by Innovate UK (grant ref: 105175) and the University of Huddersfield (grant ref: URF 510-01).

REFERENCES

[1] Dolby, “Dolby Atmos”, https://www.dolby.com/technologies/dolby-atmos, accessed 1 Dec 2020.

[2] Auro Technologies, “Auro-3D”, https://www.auro-3d.com, accessed 1 Dec 2020.

[3] DTS, “DTS:X”, https://dts.com/dtsx, accessed 1 Dec 2020.

[4] ITU-R, “Report ITU-R BS.2159-8 Multichannel Sound Technology in Home and Broadcasting Applications,” International Telecommunications Union (2019).

[5] Sony, “360 Reality Audio”, https://www.sony.co.uk/electronics/360-reality-audio, accessed 1 Dec 2020.

[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, pp. 821–830 (2014 Dec.). doi: 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, 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, pp. 870–884 (2014 Dec.).

[9] M. Lindberg, “3D Recording with 2L-Cube”, http://www.2l.no/artikler/2L-VDT.pdf, accessed on 20 June 2020.

[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, “The Psychoacoustic Testing of the 3D Multiformat Microphone Array Design, and the Basic Isosceles Triangle Structure of the Array and the Loudspeaker Reproduction Configuration,” presented at *the 134th Convention of the Audio Engineering Society* (2013 May), convention paper 8839.

[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), e-Brief 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./Feb.)
DOI: https://doi.org/10.17743/jaes.2018.0068

[17] T. Kamekawa and A. Marui, “Evaluation of Recording Techniques for Three-Dimensional Audio Recordings: Comparison of Listening Impressions Based on Difference between Listening Positions and Three Recording Techniques,” *Acoust. Sci & Tech.*, vol. 41, pp. 260-268 (2020).

[18] K. 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.

[19] H. Lee, “Multichannel 3D Microphone Arrays: A Review,” accepted for publication in *J. Audio Eng. Soc.*, vol. 69, no. 1/2, (2021 Jan./Feb.)

[20] mh acoustics, “Eigenmike microphone”, https://mhacoustics.com/sites/default/files/ReleaseNotes.pdf, accessed 1 Dec 2020.

[21] Sennheiser, “Ambeo VR Mic”, https://en-uk.sennheiser.com/microphone-3d-audio-ambeo-vr-mic, accessed 1 Dec 2020.

[22] RØDE, “NT-SF1”, https://en.rode.com/ntsf1, accessed 1 Dec 2020.

[23] Zylia, “Zylia ZM-1 Microphone”, https://www.zylia.co/zylia-zm-1-microphone.html, accessed 20 June 2020.

[24] Rumsey et al. “On the Relative Importance of Spatial and Timbral Fidelities in Judgments of degraded Multichannel Audio Quality,” *J. Acoust. Soc. Am.,* Vol. 118, pp. 968–976 (2005).

[25] R. Conetta, T. Brookes, F. Rumsey, S. Zielinski, M. Dewhirst, P. Jackson, S. Bech, D. Meares, and S. George, “Spatial Audio Quality Perception (Part 2): A Linear Regression Model,” *J. Audio Eng. Soc.*, vol. 60, pp. 847–860 (2012).

[26] S. George, “Feature Extraction for the Prediction of Multichannel Spatial Audio Fidelity,” *IEEE/ACM TALSP*, pp.1994–2005 (2006).

[27] R. Kassier, H. Lee, T. Brookes and F. Rumsey, “An Informal Comparison between Surround-Sound Microphone Techniques,” presented at *the 118th Convention of the Audio Engineering Society* (2005 May), convention paper 6429.

[28] 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, engineering brief 543.

[29] H. Lee and D. Johnson, “3D Microphone Array Recording Comparison (3D-MARCo),” Zenodo, 2019, doi: 10.5281/zenodo.3474285.

[30] International Telecommunication Union, “Advanced sound system for programme production,” *ITU-R Recomm. BS.2051-2*, 2018.

[31] G. Theile, “Natural 5.1 Music Recording Based on Psychoacoustic Principles,” *Proceedings of the AES 19th International Conference: Surround Sound Techniques, Technology, and Perception* (2001 June).

[32] H. Wittek, “Image Assistant”, https://www.hauptmikrofon.de/stereo-surround/image-assistant, accessed 20 Dec 2020.

[33] H. Lee, “A new multichannel microphone technique for effective perspective control,” *presented at the 130th Convention of the Audio Engineering Society*, 2011, convention paper 8337.

[34] R. Wallis and H. Lee, “The Effect of Interchannel Time Difference on Localization in Vertical Stereophony,” *J. Audio Eng. Soc.*, vol. 63, pp. 767–776 (2015 Oct.). doi: https://doi.org/10.17743/jaes.2015.0069

[35] R. Wallis and H. Lee, “The Reduction of Vertical Interchannel Crosstalk: The Analysis of Localization Thresholds for Natural Sound Sources,” *Appl. Sci.*, vol. 7, pp. 278 (2017). doi: https://doi.org/10.3390/app7030278

[36] K. Hamasaki, “Reproducing Spatial Impression with Multichannel Audio,” *Proceedings of the AES 24th International Conference* (2003, Jun.).

[37] E. Benjamin, R. Lee, and A. Heller, “Is My Decoder Ambisonic?,” presented at *the 125th Convention of Audio Engineering Society*, 2008, convention paper 7553.

[38] F. Zotter and M. Frank, “All-round ambisonic panning and decoding,” *J. Audio Eng. Soc.*, vol. 60, no. 10, pp. 807–820, 2012, doi: https://doi.org/10.1099/ijs.0.63370-0.

[39] F. Zotter, H. Pomberger, and M. Noisternig, “Energy-preserving ambisonic decoding,” *Acta Acust. united with Acust.*, vol. 98, no. 1, pp. 37–47, Jan. 2012, doi: 10.3813/AAA.918490.

[40] 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.

[41] D. Johnson 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), e-Brief 190.

[42] mh acoustics, https://mhacoustics.com/download, accessed 1 Dec 2020.

[43] IEM, IEM Plug-in Suite, https://plugins.iem.at/, accessed 1 Dec 2020.

[44] Aalto University, Spatial Audio Real-time Applications (SPARTA), http://research.spa.aalto.fi/projects/sparta_vsts/, accessed 1 Dec 2020.

[45] 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, pp. 2029 (2018). DOI: http://doi.org/10.3390/app8112029

[46] H. 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, convention paper 6374.

[47] J. Blauert, “On the Lag of Lateralisation Caused by Interaural Time and Intensity Differences,” *Int. J. Audiol.,* vol. 11, pp. 265-270 (1972). DOI: http://doi.org/10.3109/00206097209072591

[48] W. Grantham and F. Whiteman, “Detectability of Varying Interaural Temporal Differences,” *J. Acoust. Soc. Am.*, vol. 63, pp. 511-523 (1978). DOI: http://doi.org/10.1121/1.381751

[49] D. Griesinger, “IALF – binaural measures of spatial impression and running reverberance,” Presented at the *92nd Convention of the Audio Engineering Society* (1992), convention paper 3292.

[50] V. Hansen and G. Munch, “Making Recordings for Simulation Tests in the Archimedes Project,” *J. Audio Eng. Soc.*, vol. 39, pp. 768–774 (1991 Oct.).

[51] P. Søndergaard and P. Majdak, “The Auditory Modeling Toolbox,” in The Technology of Binaural Listening, edited by J. Blauert (Springer, Berlin, Heidelberg, 2013).

[52] L. R. Bernstein and C. Trahiotis, “The Normalized Correlation: Accounting for Binaural Detection across Center Frequency,” J. Acoust. Soc. Am., vol. 100, no. 5, pp. 3774–3784 (1996). DOI: https://doi.org/10.1121/1.417237

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

[54] 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, pp. 1637–1647 (1992).

[55] F. Zotter and M. Frank, “Efficient phantom source widening,” *Archives Acoust.,* vol. 38, pp. 27–37 (2013).

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

[57] British Standards, “Acoustics — Measurement of room acoustic parameters. Part 1: Performance spaces (ISO 3382-1:2009), 2009.

[58] T. Hidaka, L. Beranek, and T. Okano “Interaural Cross-Correlation Lateral Fraction, and Lowand High- Frequency Sound Levels as Measures of Acoustical Quality in Concert Halls,” J. Acoust. Soc. Am., vol. 98, pp. 988-1007 (1995).

[59] A. Kolarik, B. C. J. Moore, P. Zahorik, S. Cirstea, S. Pardhan, “Auditory distance perception in humans: a review of cues, development, neuronal bases, and effects of sensory loss,” Atten Percept Psychophys, vol. 78, pp. 373-395 (2016).

[60] P. Zahorik, P. “Direct-to-reverberant energy ratio sensitivity,” J. Acoust. Soc. Am., vol. 112, pp. 2110–2117 (2002).

[61] R. Y. Litovsky, B. Rakerd, T. C. T. Yin and W. M. Hartmann, “Psychophysical and Physiological Evidence for a Precedence Effect in the Median Sagittal Plane,” *J. Neurophysiol.*, vol. 77, pp. 2223–2226 (1997 Apr.).

[62] F. Zotter and M. Frank, *Ambisonics* (Springer, 2019).

[63] G. Theile, *On the Localisation of Superimposed Soundfield*, PhD Thesis (Technische Universität Berlin, 1980).

[64] S. P. Lipshitz, “Stereo Microphone Techniques: Are the Purists Wrong?,” *J. Audio Eng. Soc.*, vol. 34, pp. 717–743 (1986 Sep.).

[65] 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 July/Aug.).
doi: https://doi.org/10.17743/jaes.2018.0040

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

[67] 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, pp. 450–461 (2008).

APPENDIX

Table A. Microphone arrays included in the 3D-MARCo database. *d* and ∠ denote distance and subtended angle between microphones, respectively. See Table 1 for channel-loudspeaker mapping. Base means the base point of the microphone array (see Fig. 2).

Mic Array	Ch	Mic	Polar pattern	Configuration

PCMA-3D	FL
	FR
	FC
	RL
	RR
		DPA 4011	Cardioid	*d*(FC-Base) = 0.25m;
				d(FL-FR, RL-RR, FL(FR)-RL(RR)) = 1m;
				∠(FL(FR)-FC) = 30°; ∠((RL(RR)-FL(FR)) = 150°;
				∠(FL(FR)-Base) = -30°; ∠(RL(RR)-Base) = 0°

	FLh
	FRh
	RLh
	RRh
		DPA 4018	Supercardioid	*d*(FLh-FRh, RLh-RRh, FRh-RRh) = 1m;
				d(height layer-base layer) = 0m;
				∠(FLh(FRh)-FL(FR)) = 120° (i.e., FLh directly upwards with FL 30° tilted downwards);
				∠(RLh(RRh)-RL(RR)) = 90°

OCT-3D	FL
	FR
		DPA 4018	Supercardioid	*d*(FC-Base)=0.08m; *d*(FL-FR)=0.7m;
				d(RL-RR)=1m; *d*(FL(FR)-RL(RR))=0.4m;
				∠(FL(FR)-FC) = 90°; ∠(RL(RR)-FL(FR)) = 90°

	FC RL RR	DPA 4011	Cardioid	

	FLh
	FRh
	RLh
	RRh
		DPA 4018	Supercardioid	*d*(FLh-FRh, RLh-RRh, FRh-RRh) = 1m;
				d(height layer-base layer) = 1m;
				∠(FLh(FRh)-FL(FR)) = 120°;
				∠(RLh(RRh)-RL(RR)) = 90°

2L-Cube	FL
	FR
	FC
	RL
	RR
		DPA 4006	Omni	*d*(FC-Base) = 0.25m;
				d(FL-FR, RL-RR, FR-RR) = 1m;
				∠((FL)FR-FC) = 30°; ∠(RL(RR)-FL(FR)) = 150°;

	FLh
	FRh
	RLh
	RRh
				d(FLh-FRh, RLh-RRh, FRh-RRh) = 1m;
				d(height layer-base layer) = 1m;
				∠(FLh(FRh)-FL(FR)) = 120°;
				∠(RLh(RRh)-RL(RR)) = 90°

Decca Cuboid	FL
	FR
	FC
	RL
	RR
		DPA 4006	Omni	*d*(FC-Base) = 0.25m;
				d(FL-FR, RL-RR, FR-RR) = 2m;
				∠(FL(FR)-FC) = 30°; ∠(RL(RR)-FL(FR)) = 150°;

	FLh
	FRh
	RLh
	RRh
				d(FLh-FRh, RLh-RRh, FRh-RRh) = 2m;
				d(height layer-base layer) = 1m;
				∠(FLh(FRh)-FL(FR)) = 120°;
				∠(RLh(RRh)-RL(RR)) = 90°

Hamasaki Square (HS)	FL FR RL RR	Schoeps CCM8	Fig-of-8	*d*(FL-FR, RL-RR, FR-RR) = 2m; ∠(FL(FR)-centre line) = 90°; ∠(RL(RR)-centre line) = 90°; ∠(RL(RR)-FL(FR)) = 0°

HS height layer at 0m	FL
	FR
	RL
	RR
		DPA 4011	Cardioid	*d*(FLh-FRh, RLh-RRh, FRh-RRh) = 2m;
				d(height layer-base layer) = 0m;
				∠(FLh-FL, RLh-RL) = 135°
				(i.e., facing away from the source)

HS height layer at 1m	FL
	FR
	RL
	RR
		DPA 4011	Cardioid	*d*(FLh-FRh, RLh-RRh, FRh-RRh) = 2m;
				d(FR-FRh, RR-RRh) = 1m;
				∠(FLh-FL, RLh-RL) = 135°

Spherical (HOA)	N/A	mhAcoustics Eigenmike EM32	Raw (A-format)	See [20]