PREPRINT doi: 10.5281/zenodo.4018010

3D Microphone Array Comparison (3D MARCo): An Open-Access Database and Objective Analyses

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)

ABSTRACT

This paper presents open-access 3D sound recordings of musical performances and room impulse responses made using various 3D microphone arrays simultaneously. The microphone arrays include OCT-3D, 2L-Cube, PCMA-3D, Decca Cuboid, Hamasaki Square with height, First-order and Higher-order Ambisonics microphone systems, providing 256 possible front-rear-height combinations. The sound sources recorded were string quartet, piano trio, piano solo, organ, clarinet solo, vocal group and room impulse responses of a virtual ensemble with 13 source positions captured by all of the microphones. The recordings can be freely downloaded from https://doi.org/10.5281/zenodo.3474285. This paper also objectively analyses the microphone arrays to gain insights into spatial and tonal differences among the arrays, which will serve basis for formal subjective comparison to be conducted in the future.

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], 22.2 [4] and Sony 360 Reality Audio [5], along with the recently standardised MPEG-H [6] 3D audio codec, are being deployed widely in consumer products and streaming or broadcasting services. This new development calls for the need of new techniques and tools for 3D audio content creation. In the context of acoustic recording, over the recent years, a number of spaced 3D microphone array techniques have been proposed [7-18]. Furthermore, with the burgeoning interest in head-tracked binaural audio for virtual reality applications, Ambisonics technology [19] and Ambisonic/spherical microphone systems [20-23] came to be more widely recognised and used by practitioners than in the past. Most of the spaced 3D microphone arrays designed for 9-channel or 8-channel reproduction, although some techniques were proposed for 22.2, e.g., [12,17]. They augment existing 5-channel or 4-channel spaced microphone arrays, which are designed based on psychoacoustic principles, with additional four microphones to feed the height channels. On the other hand, the Ambisonics microphones are essentially coincident arrays and attempt to reconstruct the sound field of the recording space at the microphone position.

As more acoustic music recordings are produced in 3D nowadays, there arises the need for evaluating the qualities of such recordings both subjectively and objectively. To this end, it is first necessary to understand what kind of differences can be perceived among various types of microphone techniques, prior to rating them. There have been a number of studies that compared different 3D microphone techniques [18,24-28]. They generally showed that different techniques had different pros and cons depending on the tested attributes. However, they all used different experimental methods and had limitations in terms of the number of techniques tested, the types of sound sources used and consistency in the microphone models used for different arrays.

For a more systematic and comprehensive investigation into the perceptual characteristics of 3D microphone techniques, it would be necessary to create various types of sound sources recorded

using a number of different microphones arrays simultaneously. Furthermore, the microphones and preamps to be used should ideally be of the same brand to minimise the influence of recording systems, allowing a more controlled comparison on microphone-array-dependent spatial and timbral differences. To produce such a database of 3D recordings, the present project conducted a large-scale recording session in a concert hall using a total of 64 individual microphones, 51 of which were of the DPA d:dicate series, as well as mh acoustics Eigenmike EM32 [18] spherical microphone and Sennheiser Ambeo VR Mic [19] first-order Ambisonic microphone. Using the individual microphones, six different 9-channel or 8-channel spaced microphone arrays, comprising OCT-3D [7], PCMA-3D [8], 2L Cube [9], Decca Cuboid, and two variations of Hamasaki Square with height [10], were configured. Additional microphones for side, side height, overhead and floor channels were used with a possible extension to a larger format such as 11.0, 13.0 or 22.0 in mind. Five different types of musical performances, comprising string quartet, piano trio, organ and a cappella singers were recorded using all of the microphones simultaneously. Additionally, impulse responses of the microphones were captured for thirteen different sound source positions arranged in a typical string ensemble layout to allow for acoustic analyses of the microphone arrays as well as the creation of virtual sound sources for future experiments.

Another motivation of this project was to provide useful learning resources for spatial audio education. With a rising interest in 3D audio nowadays, there is much debate about the pros and cons of different 3D microphone techniques in the audio community, but there is a limited amount of recording resources publicly available to students. It would be ideal for students and educators to experiment with various types of techniques themselves, but this may not be practically difficult due to a large number of microphones required for a simultaneous comparison.

From the above background, the recordings and impulse responses produced in this project are provided as an open-access database [29], which is named “3D Microphone Array Comparison (3D MARCo)”. The aims of this paper are (i) to categorise and overview the microphone arrays included in the database (Sec. 1), (ii) to provide the technical details of the recording session (Sec. 2), and

(iii) objectively analyse the recordings to gain insights into physical differences among the arrays (Sec. 3). The objective data would also serve as references for subjective elicitation and grading experiments planned for the future.

1 3D Microphone Arrays included in the Database

Table 1 lists the microphone arrays included in this project, which were chosen for their distinct differences in design concepts, physical configurations and purposes. Fig. 1 illustrates their physical configurations.

Table 1. 3D microphone arrays included in the 3D-MARCo database

	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 (HOA) Ambeo VR Mic (FOA)
Ambience array	Hamasaki Square (HS) with height layer at 1m above	Hamasaki Square (HS) with height layer at 0m	

The arrays can firstly be grouped into perceptually motivated and physically motivated design concepts. The former category typically aims to achieve certain perceived characteristics in phantom imaging and spatial impression by manipulating interchannel level and time differences, whereas the latter aims to reconstruct the physical sound field in reproduction.

3D microphone arrays for music recording can be categorised into three groups according their physical configurations [30]: (i) horizontally and vertically spaced (HVS), (ii) horizontally spaced and vertically coincident (HSVC) and (iii) horizontally and vertically coincident (HVC). A wider horizontal microphone spacing tends to produce a more spacious and enveloping sound [31, 32]. However, vertical microphone spacing was found to have a minimal or no effect on overall spatial impression in 3D reproduction [8], based on which a microphone array could be made in the HSVC style, e.g.,

[8,14,17]. The perceptually motivated arrays can be further classified into “main” and “ambience” microphone arrays. The main array is defined as an array that attempts to capture both direct and ambient sounds, typically placed within the critical distance from the sound source, whereas the ambience array is dedicated to capture ambience rather than direct sound, thus being placed beyond the critical distance or using directional microphones facing away from the sound source.

Due to a limited number of microphones available, it was not possible to include all existing 3D main microphone arrays. However, it is considered that the eight array configurations and additional ambience microphones used in this project represent typical characteristics each category classified in Table 1. Even though each of the arrays was designed based on a specific philosophy as will be discussed in the following sub-sections, their front, rear, front height and rear height segments could be considered separately and combined interchangeably for creating various different configurations. For example, it is possible to derive 256 different combinations from the four segments of each of the four spaced main arrays.

The physically motivated arrays typically have a horizontally and vertically coincident (HVC) or spherical configuration. For example, first-order Ambisonic (FOA) microphones have four cardioid or subcardioid capsules arranged in tetrahedron, whereas higher-order Ambisonic (HOA) microphone systems tend to consist of multiple small capsules arranged on a small sphere. For FOA, it is also possible to natively derive Ambisonic “B-format” signals by using individual cardiod or/and figure-of-eight microphones [18,19].

The following sub-sections discusses the design principle of each of the microphone arrays. The channel names and abbreviations used in this paper and the database are as follows.

- Base layer: Front Left (FL), Front Right (FR), Front Centre (FC), Rear Left (RL), Rear Right (RR)
- Height layer: Front Left height (FLh), Front Right height (FRh), Rear Left height (RLh) and Rear Right height (RRh).

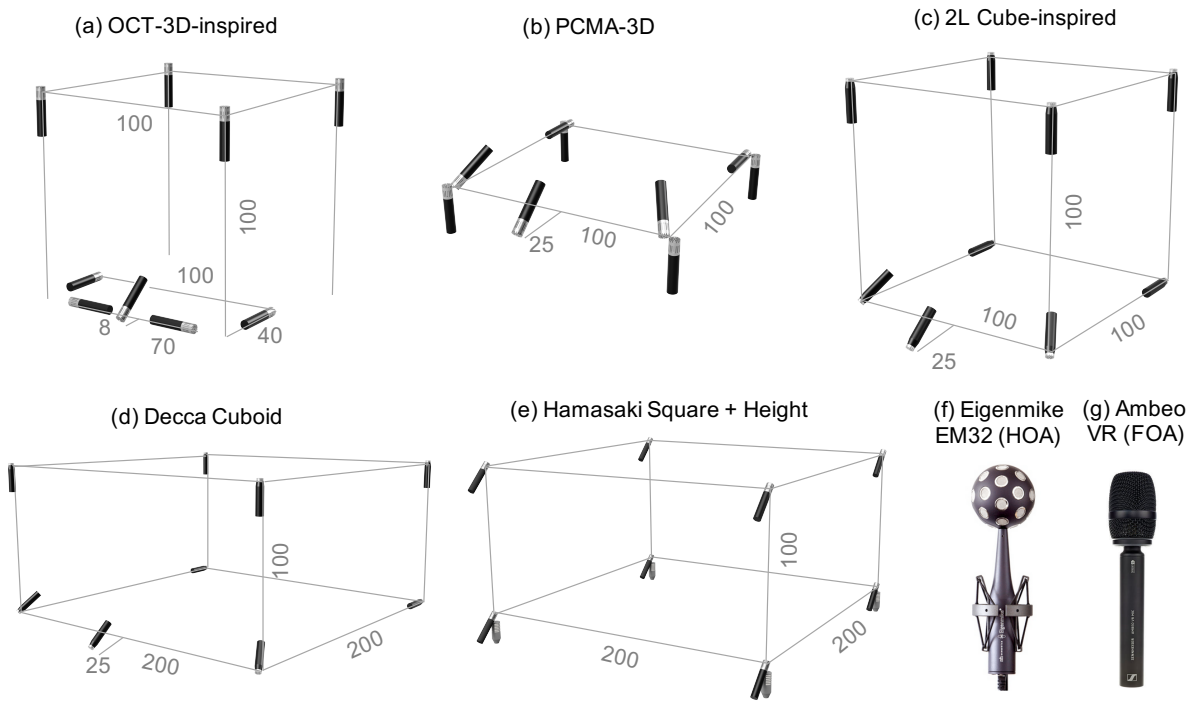


Fig. 1. Microphone arrays included in the 3D-MARCo database. Unit for the numbers is cm. Detailed information on the polar patterns and microphone angles for each array can be found in Appendix A.

1.1 OCT-3D

The OCT-3D, proposed by Theile and Wittek [7], augments the OCT (Optimised Cardioid Triangle) 5-channel main microphone array [33] with four upward-facing height microphones placed 1m above the main array (Fig. 1(a)). The OCT's front triplet uses a cardioid centre microphone placed 8cm in front the array base point and two sideward-facing supercardioid microphones of which the spacing can vary depending on the desired stereophonic recording angle (SRA)[1]. The FL-FR microphone spacing used for the current project was 70cm (i.e. OCT70). According to the 'Image Assistant' model [34,35], this produces the SRA of 115° at 2.6m source-array distance, which was the case in the current project. The rear microphones of the OCT are backward-facing cardioid microphones placed 40cm behind the FL and FR microphones, with their spacing being 20cm wider than the FL-FR spacing; e.g. 90cm RL-RR for 70cm. In the current project, however, the RL-RR spacing was chosen to be 1m for consistency with PCMA-3D, which had backward-facing RL and RR with 1m

[1] Stereophonic recording angle (SRA) is the horizontal span of the sound field in front of the microphone array that will be reproduced in full width between two loudspeakers.

spacing. PCMA-3D had 1m front-back depth as oppose to the 40cm depth of OCT-3D. This allows a comparison on the effect of array depth while the width is kept constant.

The height layer of OCT-3D originally consists of four directly upward-facing supercardioid microphones, which are placed directly above the base layer microphones apart from FC. However, it was decided to modify the height layer to be a 1m x 1m square to be consistent with the PCMA-3D's height layer. It is suggested in [36] that the vertical layer spacing of OCT-3D can be chosen between 0m and 1m. In this project, the OCT-3D's height layer was placed at 1m above the base layer to allow a comparison against the 0m vertical spacing of PCMA-3D. As mentioned earlier, the base and height layers of all arrays can be interchangeable for experimentation. Therefore, the 0m and 1m height layer spacings can be compared over any base array.

The main design aim of the OCT is to achieve accurate frontal image localisation between the front left, centre and right channels by minimising the amount of interchannel crosstalk (ICXT). The implicit assumption here is that signals from microphones other than the pair that is primarily responsible for phantom imaging is treated as unwanted crosstalk [33]. For example, if a sound source is located in the right recording sector in front of the array, signals from the microphone pair of FC and FR, which cover the recording sector where the source lies, would be considered to be 'wanted', whereas any signal from the contralateral microphone FL would be regarded as 'unwanted' crosstalk. The reduction of ICXT with OCT is realised by using sideward-facing supercardioids microphones, which have a substantially greater suppression of sounds arriving from around 90° to 135° from the on-axis compared to cardioid.

1.2 PCMA-3D

PCMA-3D is based on the PCMA (Perspective Control Microphone Array) design concept proposed by Lee [36]. The original PCMA proposal that was aimed for 5-channel surround recording. Each point in a 3-channel spaced array employs a coincident pair of forward- and backward-facing cardioid microphones to be mixed. Similarly, each rear channel uses a coincident pair of cardioid and

supercardioid microphones facing backward and sideward, respectively. By changing the mixing ratio, a virtual microphone with different direction, polar pattern and direct-to-reverberant ratio can be created, thus controlling listening perspective.

This concept has been adapted for 3D recording later based on two 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] and (ii) vertical interchannel time difference is an unstable cue for vertical phantom imaging [37]. The height microphones FLh, FRh, RLh and RRh of PCMA-3D are configured coincidentally with their corresponding base microphones FL, FR, RL and RR, as can be seen in Fig. 1(b). This forms a 'spaced XY and coincident Z' array, which would be more compact and easier to configure than a spaced XYZ array. This design concept also has a tonal benefit in 3D-to-2D downmix since there is no comb-filtering when the height microphone signals are combined with their corresponding base microphone signals.

The spacing between FL and FR is 1m, with FC 25cm in front of the base point. The subtended angle for the FL-FC and FR-FC pairs is 30°, which was chosen for the following reason. The 30° subtended angle between FL (or FR) and FC of PCMA-3D would introduce a stronger ICXT compared to the 90° angle of OCT-3D. However, as found in [32, 38], subjective preference for the presence of ICXT tend to depend on the type of sound source, and ASW or spaciousness is one of the main reasons why recordings with a stronger ICXT are preferred despite a decrease in locatedness (i.e. ease of localisation). The relatively small interchannel level differences (ICLDs) among FL, FC and FR due to ICXT are compensated by sufficient interchannel time differences (ICTDs) owing to the FL-FR spacing of 1m. The resulting localisation characteristics of the PCMA front triplet is linear, with the SRA of 106° at 2.6m source-array distance [35].

The height microphones are supercardioids and face directly upwards. This ensures that the level of the direct sound in each height layer microphone is at least 9.5 dB lower than that of the base layer, which is the minimum vertical interchannel level difference required to prevent an unwanted

elevation of the source image [39]. The sufficient vertical channel separation allows the height layer to mainly capture ambience arriving from the ceiling while the base layer captures the direct sound from the front and ambience from the back. This would provide a flexibility in balancing the level of ambience in the height channels without affecting the source image rendered mainly in the base layer.

For the rear section of PCMA-3D, a spaced pair of backward-facing cardioid microphones is coincidentally arranged with upward-facing supercardioids. The spacing between the left and right microphone pairs as well as that between the front and rear sections of the array can be decided depending on various factors such as the desired amount of interchannel decorrelation for perceived width and depth and the direct-to-reverberant (D/R) energy ratio of the rear channel signals. For this project, a 1m x 1m square arrangement was opted to allow a direct comparison against 2L Cube, which has a 1m x 1m x 1m cube configuration with omni-directional microphones. The 1m microphone spacing is able to fully decorrelate ambient signals down to around 150 Hz according to the diffuse field coherence model by Cook et al. [40].

1.3 2L Cube

The 2L-Cube is a technique developed by Lindberg [9]. It employs nine omni-directional (omni) microphones in a 1m x 1m x 1m cube arrangement, thus mainly relying on ICTD for phantom imaging, although the exact rationale for the cube arrangement is not clear from the available reference. An omni microphone would typically have an extended low-end frequency response compared to a directional microphone, which is why it is often a more preferred choice for recording a large orchestra. However, as pointed out in [8, 39], using omni microphones for all channels would introduce a large amount of ICXT in the rear and height microphone signals, which might potentially lead to an undesired imaging issue such as horizontal and vertical image shift and tonal colouration due to the comb-filter effect. Therefore, a precaution in microphone configuration and level balancing might be required. For instance, the height microphones could be made to face directly upwards rather than pointing towards the sound source. This is based on the finding that the vertical image

shift problem due to the vertical ICXT is mainly associated with frequencies above around 4 kHz [41]. Since an omni-directional microphone in practice becomes more directional as the frequency increases, using the off-axis would reduce the amount of high frequencies in the direct sounds captured by the height microphones. Alternatively, as suggested in [41], the high frequency energy of the height microphone signals could be reduced by equalisation in the post-processing stage.

Based on the above considerations, 2L Cube was adapted by angling the omni height microphones directly upwards in the current project. The physical configuration of the base layer was made identical to that of PCMA-3D; the subtended angle of FL (FR) to FC and that of RL (RR) to FC were 30° and 180°, respectively, and the spacing between FC and the base point was 25cm, producing the SRA of 114° at 2.6m source-array distance [35]. 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.

1.4 Decca Cuboid

The Decca Tree technique is widely used for large-scale orchestral recordings (e.g., a de-facto standard for film scoring). It employs three widely spaced omni microphones (FL to FR 2m – 2.5m, FC to base 1m – 1.5m), thus heavily relying on ICTD for phantom image localisation. Due to the large spacing that minimises correlation between the microphone signals, the array can provide a good sense of spaciousness in reproduction. However, the large spacing also causes a narrow SRA of 76° at 2.6m source-array distance [35] and a non-linear stereophonic image distribution due to a strong precedence effect. That is, any sound source located beyond ±38° in front of the array would be localised at the fully panned position of ±30°, with the image position shifting rapidly from 0 to ±30° as a source position changes from 0° to ±38° [35].

In this project, 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’. The

spacing between the FL and FR microphones was kept as 2m, but FC 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 a too strong centre image due to a large array depth.

1.5 Hamasaki Square with Height

The Hamasaki Square (HS) [31] is a popular technique for recording 4-channel diffuse ambience. Four sideward-facing figure-of-eight (fig-8) microphones are arranged in a square formation. The technique was developed to enhance the sense of listener envelopment (LEV) in 5-channel surround reproduction. The front microphone signals of HS feed the front left and right loudspeakers, while the rear signals are routed to the rear left and right loudspeakers. Using the front channels as well as the rear for ambience reproduction was shown to be more effective for LEV than just using the rear ones [32]. Since the microphones are oriented towards the side walls, with the null points facing towards the front, HS can sufficiently suppress the direct sound from the stage while picking up early reflections and reverberation from the lateral directions. The size of HS is recommended to be 2m to 3m [31], based on both subjective evaluation results and the diffuse field coherence model [40] suggesting a full decorrelation above 100 Hz with the 2m spacing.

Hamasaki and Van Baelen [10] extended the original HS for 3D ambience capture by adding a height layer of four upward-facing supercardioids in a square of the same size directly above it, forming an array in a cube shape. Their suggested distance between each adjacent microphone is again 2m to 3m. Additionally, if the overhead loudspeaker (a.k.a. Voice of God (VOG)) is used, an extra upward-facing supercardioid is placed in the middle of the height layer. A subjective evaluation that compared the 3D version against the original 2D HS showed that the 3D was preferred to 2D overall [10].

For the current project, a 2m x 2m HS base layer was used and two height layers of were added at 0m and 1m above it for comparison. As mentioned earlier, it was found in [8] that the vertical microphone spacing in the context of 3D main microphone array did not have a significant effect on

spatial impression. Therefore, it was considered that this may also be the case for 3D ambience capture. The polar pattern chosen for the height microphones for HS in this project was cardioid rather than supercardioid. The microphones were angled so that the null point at 180° faced the stage for a maximal suppression of direct sound. From the present authors' previous recording experiences, it was felt that cardioids facing towards the back of the recording space was more effective in the direct sound rejection and provided a warmer and more diffuse reverberation than the upward-facing supercardioids.

1.6 Spherical arrays

Spherical microphone arrays typically consist multiple microphone arrays mounted on the surface of a small sphere. It can be used either for a beamforming purpose, where a single or set of virtual microphones pointing towards different directions could be formed by processing the raw signals, or for Ambisonics, which attempts to reconstruct the original sound field in reproduction. For a spherical array to perform with a higher spatial resolution, a larger number of microphones is required. In order to derive Ambisonically encoded signals (i.e., B-format) from the raw signals from a spherical array (i.e., A-format), ideally the microphones need to be arranged in a perfectly coincident fashion (i.e. no gap between the capsules). However, this is not practically possible due to constraints in physical design. The small capsule distance leads to the so-called 'spatial aliasing', which gives rise to potential tonal quality degradation as well as localisation errors at high frequencies [42].

Furthermore, Ambisonic recordings typically suffer from a narrow listening area (sweet spot) and phasing issue with head movement in surround reproduction [19] – the listener's head essentially has to be fixed in one position for a correct sound field reproduction. Nevertheless, Ambisonics has a benefit as a delivery format since it is not restricted by any specific loudspeaker configuration; a decoder can be designed flexibly for an arbitrary loudspeaker configuration. With the recent rise of virtual reality (VR) and 360° audio-visual applications, Ambisonics came to be recognised as a useful format for efficient head-tracking in binaural headphone reproduction. Since the sound field captured can be rotated in all directions at the encoding stage, there is no need to dynamically update the

head-related transfer function (HRTF) of each loudspeaker position with head-tracking. Additionally, in binaural reproduction, the listener is always in the virtual sweet spot and therefore the limitations mentioned above are not of a concern, although research suggests that externalisation of an Ambisonic recording could be worse than that of a spaced array recording in binaural reproduction [28].

In this project, two types of spherical microphone systems were included: mh acoustics Eigenmike EM32 [20] and Sennheiser Ambeo VR Mic [21]. The Eigenmike consists of 32 capsules mounted in a small sphere with a radius of 4.2cm. Its raw signals can be processed to produce either higher-order Ambisonic (HOA) signals or individual virtual microphones (i.e. beamforming), both up to the 4th order (i.e. 25 spherical harmonic (or B-format) components). The Ambeo VR is a microphone that consists of four cardioid microphones arranged in a tetrahedron shape, from which the first-order Ambisonic (FOA) B-format components W, X, Y and Z. For the 3D reproduction of an Ambisonic recording, the raw microphone signals (so-called the A format) need to be first converted into B-format components, which then need to be decoded to loudspeakers of a certain configuration or directly for binaural reproduction. Various open-access VST plugins are available for Ambisonic encoding and decoding, such as IEM Plugin Suite[2] and Aalto University's SPARTA[3].

2 RECORDING SESSION

2.1 Recording Venue

The recordings were made in the St. Paul's concert hall in Huddersfield, UK. It is a converted church with the dimensions of 16m (W), 30m (L) and 13m (H). The venue has reverberant acoustics with the average RT60 of 2.1s, and yet provide a good clarity (e.g., clarity factor C80 [43]: 7.6 dB at 4m, 2.4 dB at 10m). The floor plan of the venue is shown in Fig. 2, and the interior can be seen in Fig. 3.

[2] https://plugins.iem.at/
[3] http://research.spa.aalto.fi/projects/sparta_vsts/

2.2 Sound Sources

Six different ensemble and solo musical sources were recorded. The musical ensemble and solo pieces are summarised below.

- String Quartet: Dvorak string quartet in G major op.106, performed by members of Up North Session Orchestra.
- Piano trio: Beethoven piano trio in E flat major, op. 1, no. 1, performed by members of Up North Session Orchestra.
- Piano solo: Chopin Nocturne in C sharp minor op. 27 & Chopin Mazurka in B flat op. 7, performed by Jonathan Fisher.
- Pipe organ: improvisation, performed by Jonathan Gooing.
- A cappella quintet: Amber Run's I found, performed by Alex Tune, Carolina Padro Calero, Emma Varley, Kitty Reid and Georgie Cooper.
- Anechoic single sources (male speech, cello, conga and trumpet) from [44], presented from a Genelec 8331A loudspeaker (45-37000 Hz) placed at 0°, -15°, -30°, -45°, -60° and -90° (in the right-hand side).

2.3 Equipment

A total of 102 channels of audio were recorded simultaneously using 64 individual microphones, Neumann KU100 dummy head (2 channels), Sennheiser Ambeo VR Mic (4 channels) and mhAcoustics Eigenmike EM32 (32 channels). The full list of the microphone models are provided in Appendix A. 51 of the individual microphones were of the DPA d:dicate series (4011 (cardioid), 4018 (supercardioid) and 4060 (omni)). They were used for all of the HVS and HSVC arrays from Table 1, the height layers of Hamasaki Square, an ORTF stereo pair, side ambience microphone pairs and a microphone for the VOG channel. Hamasaki Square consisted of four Schoeps CCM8 figure-of-eight microphones. Three AKG 414B-TLII microphones in the cardioid mode were used for the floor channels. Two CCM4s and two CMC6/MK4 cardioid microphones by Schoeps were used as spot

microphones for string instruments. Finally, a pair of Neumann KM184 cardioid microphones were used as piano spots.

All of the microphone signals apart from Eigenmike were amplified using the Merging Technologies AD8P microphone preamps installed in two Horus network audio interfaces. The gains of all of the microphones were measured prior to the recording and their differences were compensated within ±0.3 dB. The recordings were made at the sample rate of 96 kHz with 24 bits using the Reaper digital audio workstation (DAW), except for the Eigenmike that used its dedicated software recording at 48 kHz/24 bits.

2.3 Physical Setup

Fig. 2 illustrates the floor plan and the microphone array layouts. Some photos of the setup are also provided in Fig. 3. Detailed information about the microphone configurations (microphone model, polar pattern, angle and spacing) can be found in Appendix A.

All of the microphones for the OCT-9, PCMA-3D, 2L-Cube were mounted on the Grade Design Spacebar system that was custom-extended vertically with poles and 3D-printed joints, forming a cube structure (1m x 1m x 1m). Microphones for Decca Cuboid were placed on separate stands, apart from the FC microphone that shared with 2L-Cube. The main layer of the PCMA-3D was placed 2.7m high from the floor. Those of the 2L Cube and OCT-3D were 5-10cm above and below that of the PCMA-3D, respectively. Eigenmike and Ambeo were placed about 5cm and 10cm below OCT-3D's front centre microphone, respectively. The frontal microphones of the base layers were tilted downwards by about 30°

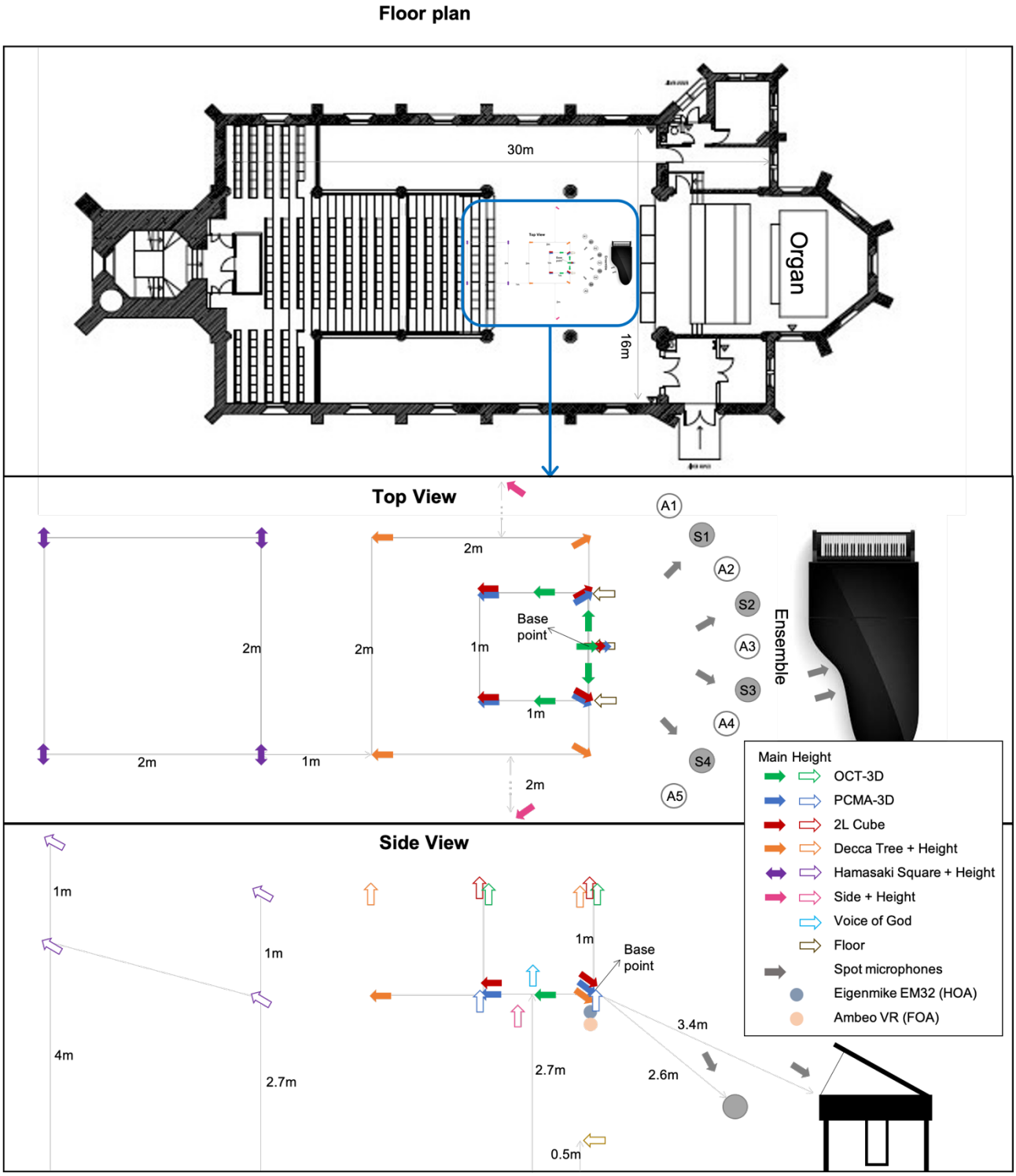


Fig. 2. Physical layout of the microphones used for the recording. S1 to S4 represent each member of the string quartet. The positions of the violin and cello in the piano trio were S2 and S3.

All of the main arrays shared the same base point. It might be argued that each microphone array should be placed at their own optimal distance from the source. Although this is true in practical recording situations, it is a subjective task involving an artistic decision. The current session kept the base point as an experimental constant since the main purpose of the project was to study perceptual

differences among the arrays at one acoustical reference point in sound field. The distance from each string instrument to the base point was 2.6m. As mentioned earlier, the SRAs of OCT-3D, PCMA-3D and 2L Cube at this distance are 115°, 106° and 114°, respectively, according to Image Assistant [35]. Therefore, having a constant base point would allow observing differences in spatial and tonal characteristics with the perceived ensemble width being kept similar.

The front pair of the Hamasaki Square was placed 3m behind the main array base point. The main array was raised at 2.7m from the floor and the 0m and 1m height layers of cardioid microphones described earlier were placed directly above the main layer microphones.

Further to the above-mentioned microphone arrays, additional microphones were placed to feed the side, side height, floor and overhead (VOG) channels for a larger reproduction format such as 22.2. The side and side height microphones were configured in a vertical coincident fashion, with sideward-facing cardioid and upward-facing supercardioids microphones. The VOG microphone was a supercardioid facing directly upwards and placed in the middle of the main array structure. Three backward-facing floor channel microphones were placed directly below the front three microphones of the PCMA-3D and 50cm above the floor.

Each member of the string quartet was placed at 45°, 15°, -15° and -45° from the bottom of the main microphone stand. Each singer of the a cappella quintet was positioned from 60° to -60° with 30° intervals. The ensembles would have a narrower angular spread in a typical concert formation, but it was decided to have the wider spread for a wider perceived ensemble width. Furthermore, the regular angular interval between each musician would allow examining the angular distribution of each corresponding phantom source in reproduction. The distance from the base point to each musician was 2.6m for the string quartet and 2.4m for the singers. The violin and cello of the piano trio were position at 15° and -15°. The piano and the pipe organ were 3.4m and 12m away from array base point, respectively.

Fig. 3. Photos of the recording venue and microphone array setup.

2.5 Microphone Array Impulse Responses

Furthermore, impulse responses for all microphones used were captured for a virtual ensemble of thirteen source positions, using the exponential sine sweep method [45] implemented within the HAART software [46]. Genelec 8331A loudspeakers were used as sources and their acoustic centre was at 1.14m above the floor. The source azimuths measured from the loudspeaker level directly below the base point ranged from -90° to 90° with 15° intervals. 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°. These microphone array impulse responses (MAIRs) are expected to be useful for creating virtual 3D recording stimuli for spatial audio research and education. They were also used for calculating various objective parameters for different arrays as described in the next section.

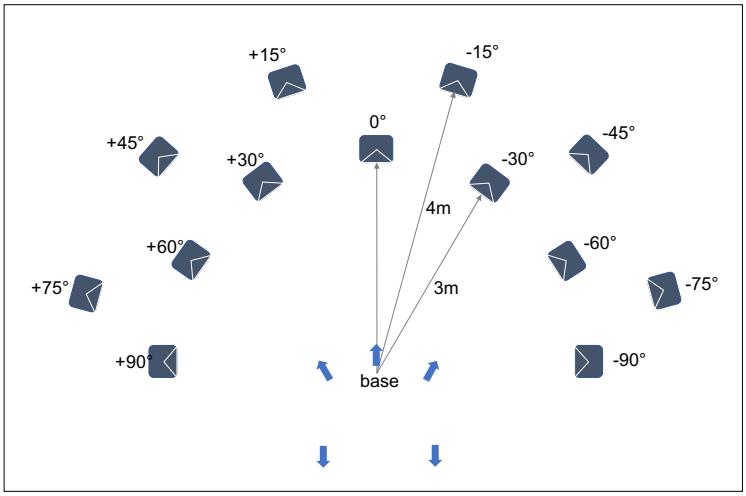


Fig. 4. layout of the virtual loudspeaker ensemble used for capturing microphone array impulse responses.

3 OBJECTIVE ANALYSES

In order to gain insights into objective differences between the microphone arrays, this section computes a set of parameters listed below, which might be associated with different types of perceptual attributes, such as horizontal and vertical image stability, tonal colouration, ASW, LEV, vertical image spread and perceived source distance.

- Interchannel level and time differences 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 cross-correlation coefficient (ICCC).
- Interaural cross-correlation coefficient (IACC).
- Direct-to-reverberant energy ratio (DRR).

The results would serve as references for hypothesising and explaining perceived differences between the arrays, which will be formally investigated in future subjective studies. This section first presents the general methods used for the analyses, followed by the detailed description of each parameter and the result and discussion in the following sub-sections.

3.1 General Method

The analysis strategy used here was adapted from [8]. Two types of signals were used for the analysis: (i) microphone array impulse responses (MAIRs) taken directly from the database and (ii) binaural impulse responses of reproduction (BIRR), which were synthesised by convolving the MAIRs with the head-related impulse responses (HRIRs) of corresponding virtual loudspeakers. The MAIRs were used for the measurements of the ICLD and ICTD of ICXT, ICCC and DRR. For the BIRR, ILD, ITD, IACC and the spectral influence of the height layer were measured. The impulse responses were used since they can be decomposed into specific time windows of direct sound, early reflections and late reverberation, allowing controlled investigations into source-related and environment-related perceptual properties of different microphone techniques.

The sound source position chosen for this investigation was +45° (see Fig. 4). This position was considered to be suitable for the purpose of this analysis for the following reasons. Considering the SRAs of OCT-3D (115°), PCMA-3D (106°) and 2L-Cube (114°) from the base point, the perceived position of a source at +45° would be shifted from the front centre to the front left loudspeaker by around 70% according to Image Assistant [35]. This means that there would be interchannel and interaural differences that are sufficient enough for one to observe the perceptual effects of the microphone configurations. For readers who are interested in performing further analysis for other source positions, necessary MATLAB codes are provided in the 3D-MARCo database [29].

Fig. 5 describes the overall workflow of the analyses, and Table 2 presents the 9-channel virtual loudspeaker configuration used. The azimuth and elevation angles of the loudspeakers were chosen based on ITU-R BS.2159-8 [47]. 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. Each MAIR of the spaced arrays is discretely routed to each corresponding loudspeaker in reproduction, and therefore there was no further processing or mixing required. On the other hand, the Eigenmike's raw signals need to go through a series of processing to obtain the loudspeaker signals. The raw signals were first converted into B-format spherical harmonics using the "EigenUnit" plugin[4], which were then decoded to the loudspeakers configured as in Table 2. The All-Round Ambisonic Decoder (ALLRAD) [48] in the IEM plugin suite was used since it allows decoding to an unevenly distributed arrangement such as the one used here. 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 plugin suite or SPARTA on the Reaper session template provided with the database.

[4] https://mhacoustics.com/download

The BIRRs were synthesised by convolving the MAIRs with the KU100 dummy head HRIRs taken from the SADIE II database [49]. The MAIRs and BIRRs went through a time-window segmentation as required for each of the parameters (described in each corresponding sub-section below). For ICCC and IACC, the segmented signals were then split into nine octave bands with their centre frequencies ranging from 63 Hz to 16 kHz, using an 8th order biquad linear-phase filter (-48 dB/oct). For ILD and ITD, the BIRRs were split into 64 equivalent rectangular bands (ERBs) through a Gammatone filter bank [50]. Half-wave rectification and a first-order low-pass filtering at 1 kHz were applied to mimic the breakdown of the phase-locking mechanism [26, 27]. ILDs and ITDs for each band were calculated for 50%-overlapping 50ms frames with Hann window. Detailed definitions and methods used for the computations of the parameters are provided in the following sections.

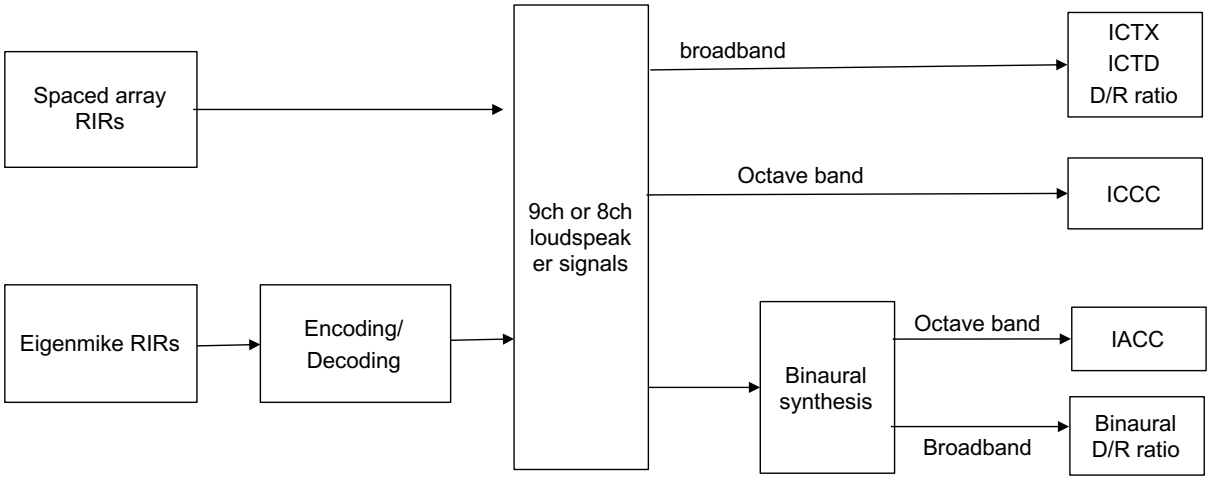


Fig. 5. Add time segmentation, spectral differences of BIRR

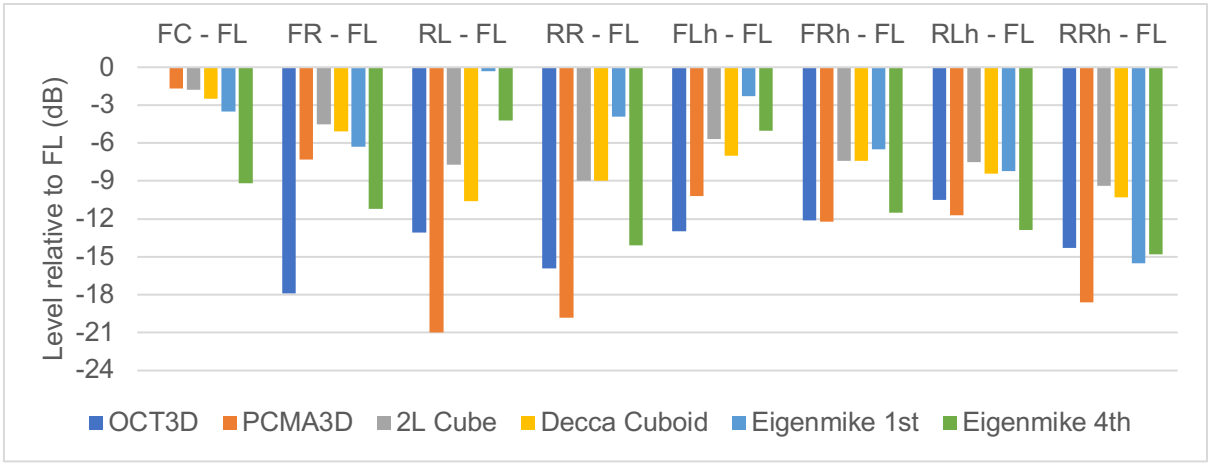
Table 2. Channel-loudspeaker configuration used for the creation of BIRR stimuli used for the objective measurements (anticlockwise angular orientation).

Channels	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

3.2 Interchannel Level and Time Differences of Crosstalk

As described in Sec. 1.1, 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. That is, 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 [38]. Between vertically oriented microphones (e.g., FL and FLh), on the other hand, ICXT would cause the phantom source to be shifted upwards regardless of ICTD [39] if it is not suppressed more than 7 dB compared to the direct sound in the base channel.

Equation? ICLD, ICTD



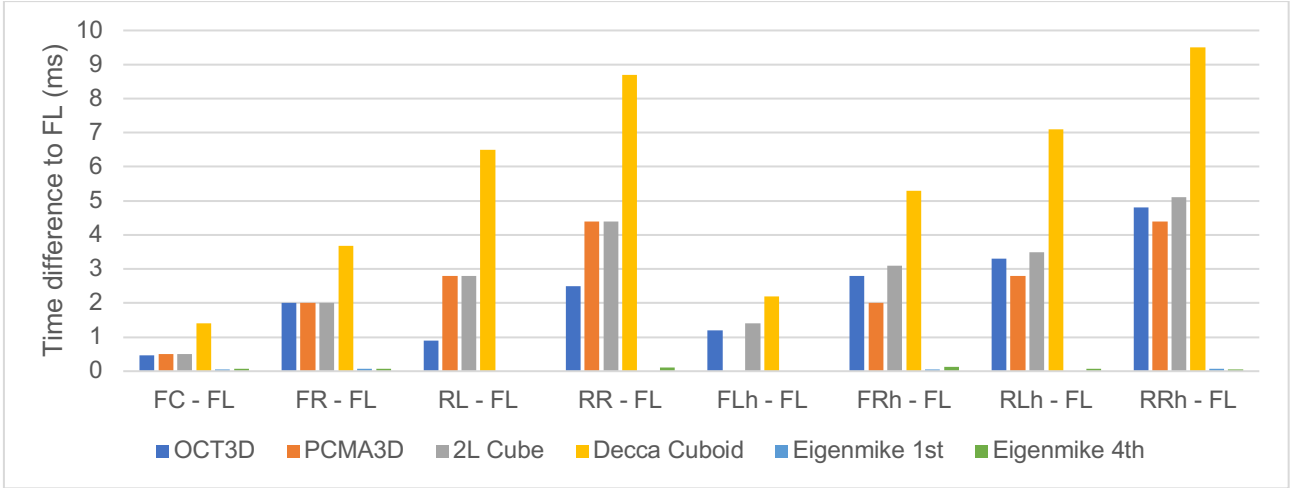


Fig. 6. Interchannel level and time differences (ICLD and ICTD) of each microphone to FL, measured using the energy of the direct sound portion (0 to 2.5ms) of the impulse responses captured for the +45° source position. ICLDs were not calculated for the Hamasaki Square arrays as their aim is to capture ambience. ICTDs were not calculated for Eigenmike since it is a coincident array.

Fig. 6 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 2.5ms after the initial impulse). 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 [33], FL and FC would be the responsible pair 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.

Firstly, looking at the horizontal channel pairs, 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 delayed 2.8ms to FL, whereas that of OCT-3D was delayed 0.9ms. 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 OCTD-3D and PCMA-3D, is large enough to trigger the precedence effect in combination with the ICLD, thus locating the phantom image in the front side, the better front-rear separation of PCMA-3D might provide more headroom for increasing the level of the rear ambience without affecting the frontal phantom image.

2L cube and Decca Cuboid generally had stronger ICXT than OCT-3D and PCMA-3D due to the use of omni-directional microphones. Nevertheless, their ICTDs to FL were larger than 1ms for all pairs, which would be sufficient to trigger the precedence effect for localisation between the horizontal

channels. However, as reported in [37], 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 7.5 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 level of ambience in the height channels of OCT-3D and PCMA-3D could be raised around 3 dB to 6 dB without affecting the localisation of the source image, whereas doing the same with 2L Cube or Decca Tree would not only cause the loudness but also shift the image upwards. Note that the height omni microphones of these arrays 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 effects.

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 [19].

3.3 Fluctuations in Interaural Level and Time Differences

The interchannel level and time differences among the microphone signals eventually produce interaural level and time differences (ILD and ITD) when they are reproduced from the loudspeakers. It is well known that the ILD and ITD cues determine the perceived horizontal position of a sound image. However, when there is an modulation between two or more signals, the ILD and ITD tend to vary over time [54,55,56] and this type of fluctuation has been found to be related to the movement of the image or the perceived spread of the image, depending on the fluctuation rate (i.e., the "localisation lag" phenomenon [54]). That is, at low rates of fluctuations (< 3 – 20 Hz, depending on the experimental method and the type of signal [54,55,56), the image would be perceived to be moving between left and right, whereas high 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 imaging stability and ASW.

Since 3D microphone array signals include a direct sound with different ICLD and ICTD relationships in every channel, when they are summed at the ears in reproduction, it can be expected that fluctuations in ILD and ITD would occur and their magnitude would differ depending on the array. To measure the fluctuations in the current study, the BIRRs for the 45° source position up to 10ms after the earliest direct sound were convolved with 10-second-long pink noise and anechoic trumpet recording [44] 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 boundary value of 10 ms was chosen to include the direct sounds from all nine channels of each array; the maximum ICTD to FL observed amongst all arrays was 9.5 ms for RRh – FL of Decca Cuboid (Fig. 6(b)). The ILDs and ITDs were first measured for each ERB in each 50%-overlapping 50ms frame. The definition of ITD (time delay of the left ear signal to the right one) used here was the lag (ms) of the maximum interaural cross-correlation function (Eq. 1) [57]. 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 ERB with the centre frequency of 1.47 kHz and below, while the ILDs were averaged for the ERBs with the centre frequencies from 1.62 kHz to 19 kHz.

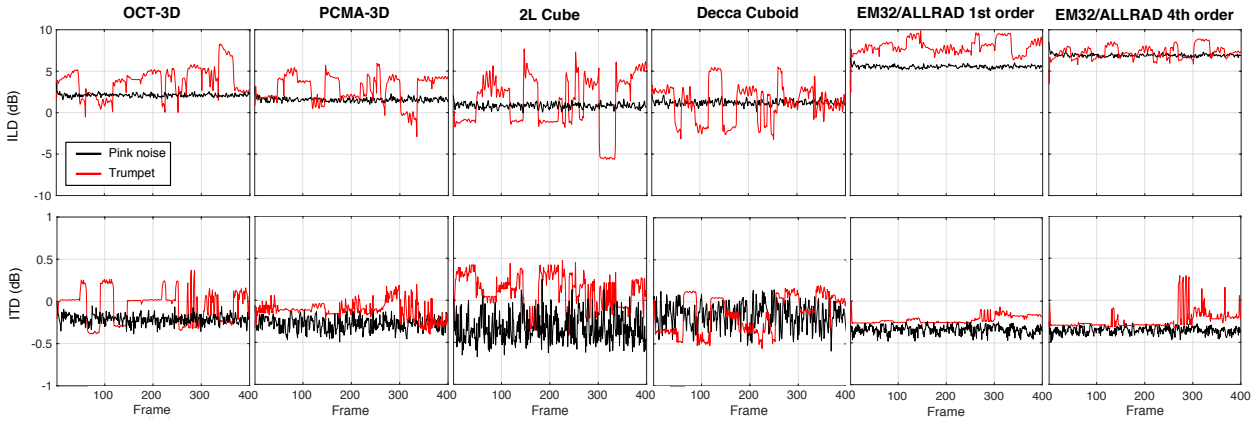


Fig. 7. ILDs and ITDs measured for the 50%-overlapping 50ms Hann-windowed frames of 10-second-long pink noise (black) and anechoic trumpet (red). The ILD and ITD for each frame are the averages of ILDs and ITDs computed for the ERBs with the centre frequencies between 1.62 kHz and 19 kHz and for those up to the centre frequency of 1.47 kHz, respectively.

In Fig. 7, 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 techniques cause a greater magnitude of ITD fluctuation over time than the coincident techniques, which is also in line with Lipshitz [57]'s observation on 2-channel stereo microphone techniques. Furthermore, considering that OCT-3D and PCMA-3D or 2L Cube had almost the same array size but they differed in the amount of ICXT, it can be also suggested that an array with a greater amount of ICXT would cause a greater magnitude of ITD fluctuation.

The differences in ITD fluctuation observed for the noise 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 [54,55,56]. 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

rate was between 21 Hz and 31 Hz, which are considered to be high enough to suggest an ASW perception without an audible image movement.

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. 7, 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, whereas the magnitude of fluctuation between each polarity flip was relatively minor. In contrast, the coincident Ambisonic conditions 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. The HSVC array PCMA-3D had a moderate fluctuation pattern, with smaller 3SD than the HVS arrays for both ILD and ITD. 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 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, as discussed in Sec. 3.2, not only the base layer but also the height layer picks up a certain amount 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.

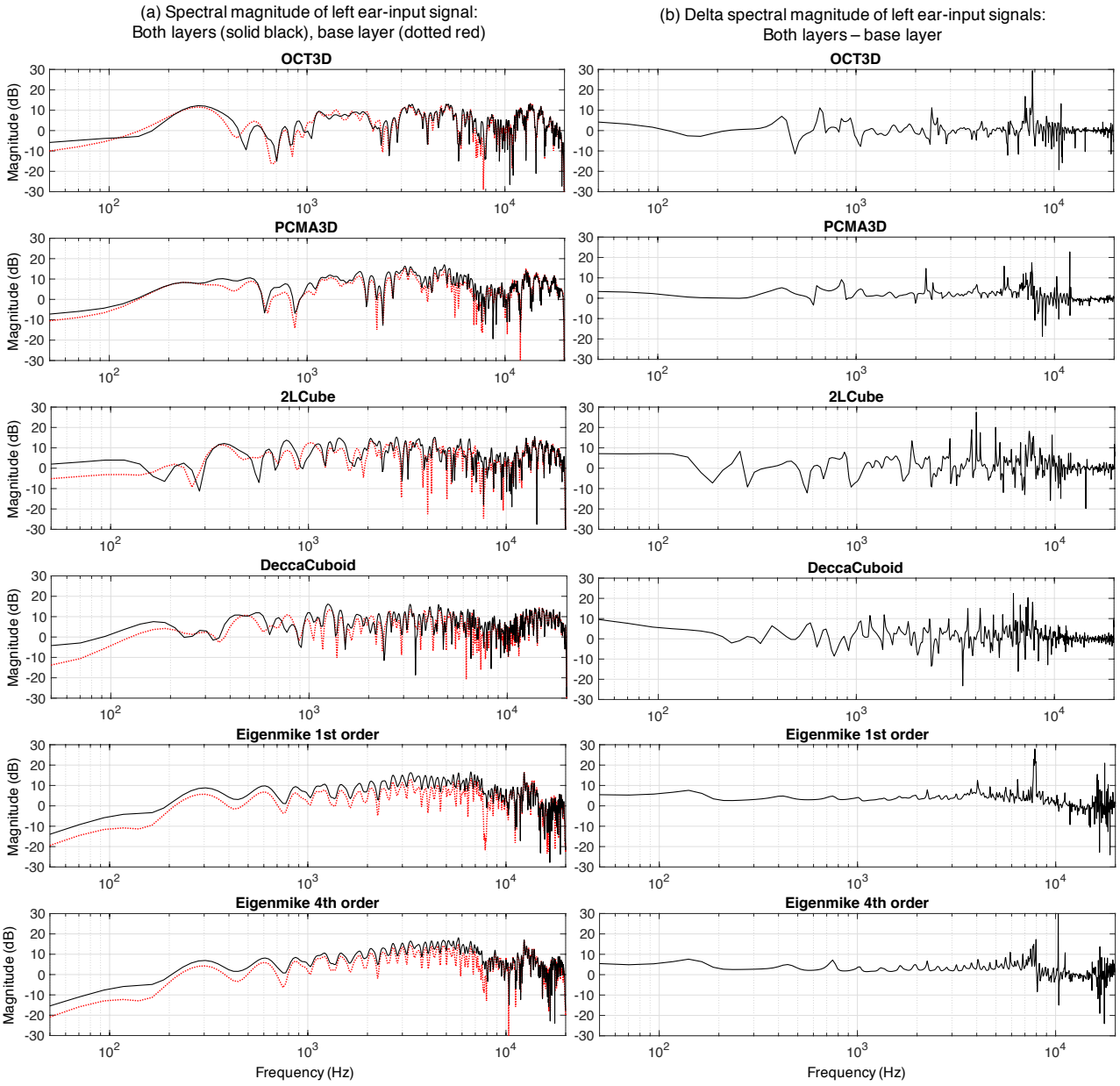


Fig. 9. Spectral magnitude difference. Say magnitude spectrum in the title

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. As in the ILD and ITD analyses above, this was done for the time window up to 10 ms after the earliest direct sound in the BIRR of

each array. The results are shown in Fig. 9. The delta plots in the right columns represent the effect of the addition of the height layer on 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. 9(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. 6. 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.

Observing the magnitude spectra of both layers in Fig. 9(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 over frequency responses were most even owing to the coincident nature, despite the comb-filter pattern due to the floor reflection at 2.8ms that was included in the time window.

The above analyses 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" [58] 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.

3.5 Interchannel Cross-Correlation Coefficient (ICCC)

Interchannel cross-correlation coefficient (ICCC) is as a useful measure of the similarity between signals and known to be associated with auditory image spread in horizontal stereophonic reproduction [59,60] and listener envelopment (LEV) [61]. It is also related to the size of listening area (i.e. sweet spot) [31].

For this investigation, the MAIRs were first segmented into early and late portions (ICCC E: t1 = 0 ms to t2 = 80 ms; ICCC L: t1 = 80 ms to t2 = 2100 ms) in order to predict differences in source-related and environment-related width attributes. The 80ms boundary point between the two segments is typically used for musical sources in concert hall research [43]. The segmented signals were then split into nine octave bands with their centre frequencies ranging from 63 Hz to 16 kHz. ICCC 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 ICCC $E(L)_{Low}$, ICCC $E(L)_{Mid}$, ICCC $E(L)_{High}$. As in ICXT, ICCCs were calculated for all channel signals against FL, which was closest to the sound source. Additionally, the other symmetrical pairs of RL-RR, FLh-FRh and RLh-RRh were included in the analysis.

Fig. 10 shows the results of the ICCC analyses. At a glance, it is apparent that the low band ICCCs 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 ICCCs for all bands. The difference between the early and late segments was also minimal for most spaced array conditions. On the other hand, the ICCCs 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.

Fig. 10. Interchannel cross-correlation coefficients for various pairs of microphone signals.

Differences between the spaced microphone arrays appear to be most obvious at the low bands. For FL-FR, the Decca Cuboid had the lowest ICCC 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 ICCC 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 ICCC. 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 ICCC L_{Low} were much smaller than those in ICCC 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 ICCC E_{Low} and ICCC 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 ICCCs 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.

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 ICCCs were comparable to those of the other spaced main arrays in general. Gribben and Lee [REF] found that in a 9-channel loudspeaker reproduction, the effect of vertical ICCC 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 ICCCs 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 ICCC.

Griesinger [61] 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 ICCC 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 ICCC 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 ICCC 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 ICCCs 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 front-back head movement would still exist even at the higher order. However, it is worth noting that the ICCCs of the Ambisonic loudspeaker signals and their dependency on the order might heavily depend on the type of decoder used. The ALLRAD used for the current analysis used the “basic” weighting, which is optimised for sound field reconstruction at frequencies below around 700 Hz. The result might be different if the decoder used the “max rE” weighting, which is optimised for the imaging of 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 ICCCs than the main arrays at the low bands, whereas the differences were negligible for most channel pairs in both segments. However, the ICCCs for the main and ambience arrays for the early segment might have different perceptual effects. Since the direct-to-reverberant (D/R) energy ratios of all of the HS array signals were much lower than those of the main array signals, the ICCC 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 ICCCs 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)

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 normalised cross-correlation function, for binaural room impulse responses (BRIRs) over the lag range of -1 ms and +1 ms. Hidaka et al. [62] 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 IACC E3 and IACC L3 for the early and late segments, respectively.

For the current analysis, IACC E3 and IACC L3 were computed for BIRRs synthesised for each of the base and height loudspeaker layers separately as well as both layers. The results are plotted in Fig. 11 (a) to (c). Additionally, Fig. X (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 ICCC 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 [63]. 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 between the base layer only and both layers 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.

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.

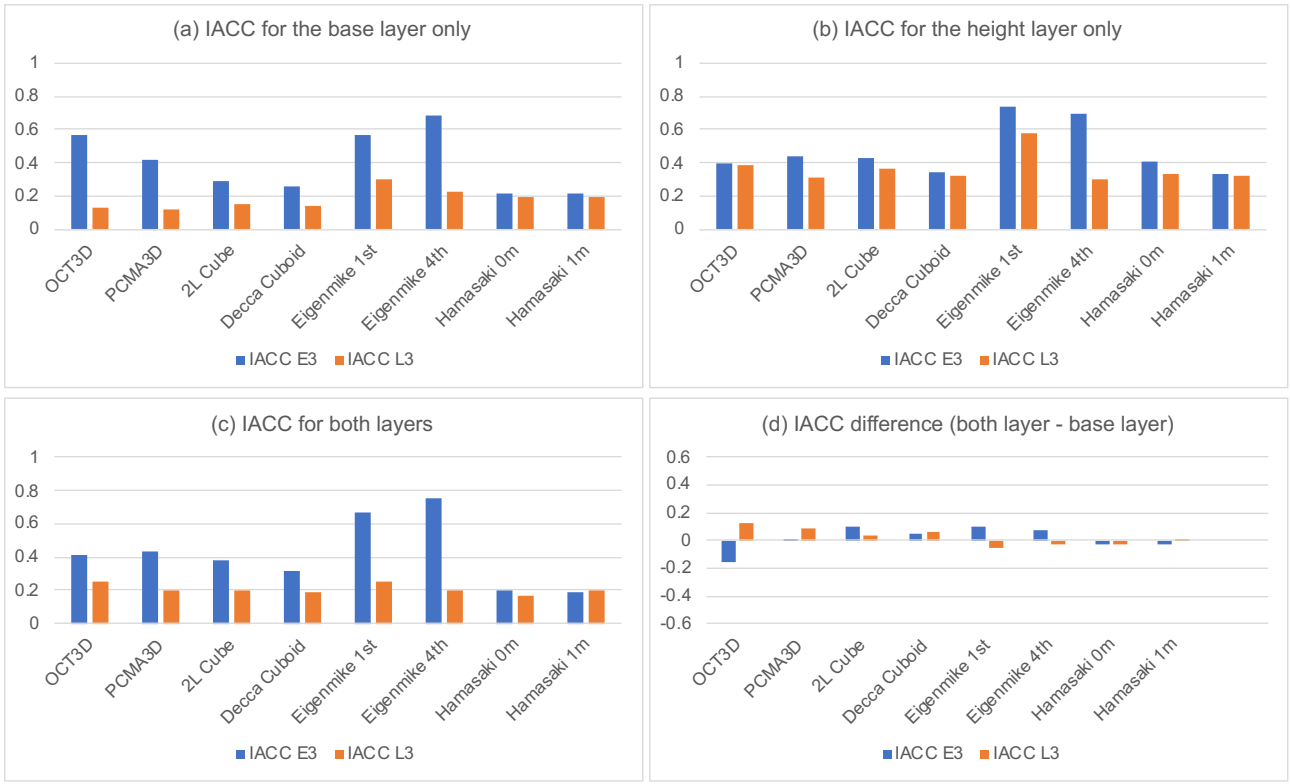


Fig. 11. Interaural cross-correlation coefficients (IACCs) for ear-input signals resulting from different microphone signals reproduced from a binaurally synthesised 9-channel 3D loudspeaker system.

3.7 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 [64]. 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 2.5ms since it is approximately the duration of anechoic HRIR and is short enough to exclude the first reflection. 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 2.5ms plus the maximum ICTD from the earliest signal (FL in the current case).

Fig. 12 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.

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 4^{th} for the left ear, and 2.7 dB between Eigenmike 4^{th} 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. [65] 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 [66] 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 D/R ratio of the binaural signals or a channel-dependent weighting of D/R ratio that affects the perceived distance. This should be clarified in a future subjective study.

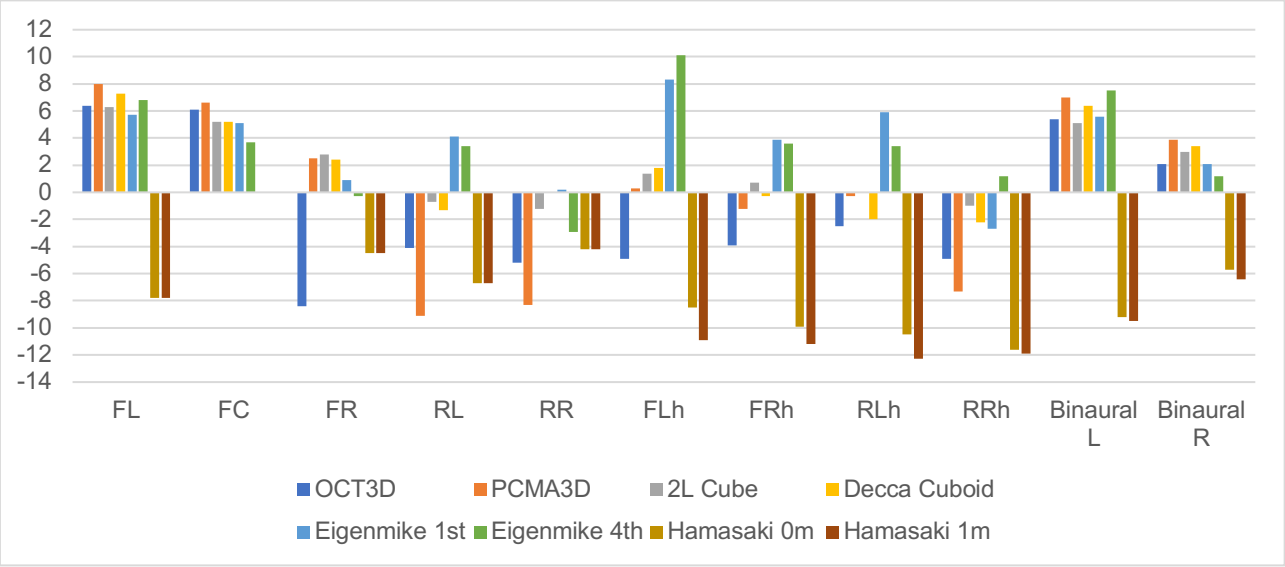


Fig. 12. Direct-to-Reverberant Ratio (DRR) for each microphone and ear-input signal.

4 SUMMARY

As the first part of a series of investigations to follow in the topic of 3D acoustic recording quality evaluation, an extensive set of sound recordings of various musical performances and room impulse responses was produced in a concert hall using eight different 3D microphone array configurations and additional microphones simultaneously. They are available as an open-access database named “3D Microphone Array Comparison (3D MARCo)” under the CC-BY NC 3.0 license (i.e., free to share and adapt the material, but not permitted to use for commercial purposes). The database will be used for future studies that will establish attribute scales and an objective model for the evaluation of 3D acoustic recording quality. It is also expected to be a useful resource for spatial audio education and critical ear training.

This paper also described the objective measurements of differences among the microphone arrays, which were conducted using various parameters that might be associated with different perceptual attributes. The aim of these analyses was to provide objective data and hypotheses to support future subjective studies to be conducted on the perceptual differences between the arrays, rather than drawing conclusions. The observations from the analyses generally suggest the following.

- There were substantial differences among the 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 a considerable difference in the perceived magnitudes of horizontal spatial impression (e.g., ASW and LEV) and the size of listening area due to the different degree of interchannel decorrelation. Considerable differences in vertical decorrelation were also observed, but based on previous research, this is hypothesised to have a minimal effect on perceived vertical image spread, based on the literature.
- 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, even though the base and height layers might have audible differences independently.
- The differences in the D/R ratios of ear-input signals resulting from the 9-channel playback were around or below the just noticeable difference of perceived auditory distance, even though individual microphones had larger differences especially in the rear channels. This raises an interesting question as to whether it would be the channel-dependent balance of 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 to establish a set of defined attribute scales, which will then be used for a rating experiment. 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. In addition, the arrays will be compared in terms of their horizontal and vertical phantom imaging accuracy.

5 ACKNOWLEDGMENT

This project was partly funded by Innovate UK (Ref: 105175). The authors would like to give special thanks to Eddy Brixen and DPA Microphones for providing the majority of microphones used for the

project, Paul Mortimer of Emerging UK and Claude Cellier of Merging Technologies for providing a Horus audio interface and AD8P microphone preamps. They are also grateful to all of the musicians from Up North Session Orchestra and the University of Huddersfield's Music Department who performed. Last but not least, the authors thank Bogdan Bacila for his technical support with the 3D-printing of microphone clamps, and other members of the Applied Psychoacoustics Lab who assisted on the project.

References

[1] Dolby, "Dolby Atmos", https://www.dolby.com/technologies/dolby-atmos, accessed 20 June 2020.

[2] Auro Technologies, "Auro-3D", https://www.auro-3d.com, accessed 20 June 2020.

[3] DTS, "DTS:X", https://dts.com/dtsx, accessed 20 June 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 20 June 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, "3D Audio Natural Recording," *Proceedings of the 27th Tonmeistertagung* (2012).

[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, “Die Kirche, das Dorf und 3D-Audio – was gehort wohin?” https://www.youtube.com/watch?v=osZ842Zaj5Q, accessed 20 June 2020.

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

[20] mh acoustics, “Eigenmike microphone”, https://mhacoustics.com/products, accessed 20 June 2020.

[21] Sennheiser, “Ambeo VR Mic”, https://en-uk.sennheiser.com/microphone-3d-audio-ambeo-vr-mic, accessed 20 June 2020.

[22] RØDE, “NT-SF1”, https://en.rode.com/ntsf1, accessed 20 June 2020.

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

[24] U. Scuda, H. Stenzel and H. Lee, “Perception of Elevated Sound Image Recorded with 3D-Audio Microphone Arrays,” *Proceedings of International Conference on Spatial Audio* (2013), pp. 122-128.

[25] W. Howie, R. King, D. Martin and F. Grond, “Subjective Evaluation of Orchestral Music Recording Techniques for Three-Dimensional Audio,” presented at *the 142nd Convention of the Audio Engineering Society* (2017 May), convention paper 9797.

[26] E. Bates, S. Dooney, M. Gorzel, H. O’dwyer, L. Ferguson, F. M. Boland, “Comparing Ambisonic Microphones – Part 2,” presented at *the 142nd Convention of the Audio Engineering Society* (2017 May), convention paper 9730.

[27] L. Riitano and M. Victoria, “Comparison between different Microphone-Arrays for 3D Audio,” presented at *the 144th Convention of the Audio Engineering Society* (2018 May), convention paper 9980.

[28] C. Millns and H. Lee, “An Investigation into Spatial Attributes of 360° Microphone Techniques for Virtual Reality,” presented at *the 144th Convention of the Audio Engineering Society* (2018 May), convention paper 10005.

[29] H. Lee and D. Johnson “3D Microphone Array Comparison (3D MARCo) (Version 1.0.1) [Data set]” https://doi.org/10.5281/zenodo.3474285, accessed 22 June 2020.

[30] H. Lee, “Multichannel 3D Microphone Array Techniques for Music and Ambience Recording: An Overview”, Under review for publication in J. Audio Eng Soc.

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

[32] 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.

[33] 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).

[34] H. Wittek and G. Theile, “The Recording Angle – Based on Localization Curves,” presented at *the 112th Convention of the Audio Engineering Society* (2002 May), convention paper 5568.

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

[36] H. Lee, “A New Microphone Technique for Effective Perspective Control,” presented at the *130th Convention of the Audio Engineering Society* (2011 May), convention paper 8337.

[37] 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

[38] H. Lee, *Effects of Interchannel Crosstalk in Multichannel Microphone Technique*, PhD Thesis (University of Surrey, 2006). http://eprints.hud.ac.uk/id/eprint/9684/

[39] 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

[40] R. K. Cook, R. V. Waterhouse, R. D. Berendt, S. Edelman, and M. C. Thompson, Jr., “Measurement of Correlation Coefficients in Reverberant Sound Fields,” *J. Acoust. Soc. Am.,* Vol. 27, pp. 1072–1077 (1955).

[41] R. Wallis and H. Lee, “Localisation of Vertical Auditory Phantom Image with Band-limited Reductions of Vertical Interchannel Crosstalk,” Appl. Sci., vol. 10, pp. 1490 (2020). DOI: https://doi.org/10.3390/app10041490

[42] E. Kurz, F. Pfahler and M. Frank, “Comparison of first-order Ambisonics microphone arrays,” *Proceedings of International Conference on Spatial Audio* (2015).

[43] BSi, “ISO 3382-1 Acoustics – Measurement of Room Acoustic Parameters – Part 1: Performance Spaces” (2009).

[44] 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.).

[45] 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.

[46] 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.

[47] ITU-R, Recommendation ITU-R BS.2159-8: Multichannel Sound Technology in Home and Broadcasting Applications (2019).

[48] F. Zotter and M. Frank, “All-Round Ambisonic Panning and Decoding,” *J. Audio Eng. Soc.*, vol. 60 (2012 Oct.).

[49] 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

[50] P. Søndergaard and P. Majdak, “The Auditory Modeling Toolbox,” in The Technology of Binaural Listening, edited by J. Blauert (Springer, Berlin, Heidelberg, 2013). DOI: https://doi.org/10.1007/978-3-642-37762-4

[51] 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

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

[53] 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).

[54] 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

[55] 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

[56] 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.

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

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

[59] G. Kendall, “The Decorrelation of Audio Si gnals and Its Impact on Spatial Imagery,” *Comput. Music. J.*, pp. 71-87 (1995).

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

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

[62] 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).

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

[64] 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).

[65] 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).

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

APPENDIX A

Table A. Microphone arrays included in the 3D-MARCo database.

<table>
<tr><th>File no.</th><th>Mic Array</th><th>Ch</th><th>Mic</th><th>Polar pattern</th><th>Configuration</th></tr>
<tr><td>1</td><td rowspan="9">PCMA-3D</td><td>FL</td><td rowspan="5">DPA 4011</td><td rowspan="5">Cardioid</td><td rowspan="5">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°</td></tr>
<tr><td>2</td><td>FR</td></tr>
<tr><td>3</td><td>FC</td></tr>
<tr><td>4</td><td>RL</td></tr>
<tr><td>5</td><td>RR</td></tr>
<tr><td>6</td><td>FLh</td><td rowspan="4">DPA 4018</td><td rowspan="4">Supercardioid</td><td rowspan="4">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°</td></tr>
<tr><td>7</td><td>FRh</td></tr>
<tr><td>8</td><td>RLh</td></tr>
<tr><td>9</td><td>RRh</td></tr>
<tr><td>10</td><td rowspan="9">OCT-3D</td><td>FL</td><td rowspan="2">DPA 4018</td><td rowspan="2">Supercardioid</td><td rowspan="5">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°</td></tr>
<tr><td>11</td><td>FR</td></tr>
<tr><td>12</td><td>FC</td><td rowspan="3">DPA 4011</td><td rowspan="3">Cardioid</td></tr>
<tr><td>13</td><td>RL</td></tr>
<tr><td>14</td><td>RR</td></tr>
<tr><td>15</td><td>FLh</td><td rowspan="4">DPA 4018</td><td rowspan="4">Supercardioid</td><td rowspan="4">d(FLh-FRh, RLh-RRh, FRh-RRh) = 1m;
d(height layer-base layer) = 1m;
∠(FLh(FRh)-FL(FR)) = 120°;
∠(RLh(RRh)-RL(RR)) = 90°</td></tr>
<tr><td>16</td><td>FRh</td></tr>
<tr><td>17</td><td>RLh</td></tr>
<tr><td>18</td><td>RRh</td></tr>
<tr><td>19</td><td rowspan="9">2L Cube</td><td>FL</td><td rowspan="9">DPA 4006</td><td rowspan="9">Omni</td><td rowspan="4">d(FC-Base) = 0.25m;
d(FL-FR, RL-RR, FR-RR) = 1m;
∠((FL)FR-FC) = 30°; ∠(RL(RR)-FL(FR)) = 150°;</td></tr>
<tr><td>20</td><td>FR</td></tr>
<tr><td>21</td><td>FC</td></tr>
<tr><td>22</td><td>RL</td></tr>
<tr><td>23</td><td>RR</td><td rowspan="5">d(FLh-FRh, RLh-RRh, FRh-RRh) = 1m;
d(height layer-base layer) = 1m;
∠(FLh(FRh)-FL(FR)) = 120°;
∠(RLh(RRh)-RL(RR)) = 90°</td></tr>
<tr><td>24</td><td>FLh</td></tr>
<tr><td>25</td><td>FRh</td></tr>
<tr><td>26</td><td>RLh</td></tr>
<tr><td>27</td><td>RRh</td></tr>
<tr><td>28</td><td rowspan="2">Decca Cuboid</td><td>FL</td><td rowspan="2">DPA 4006</td><td rowspan="2">Omni</td><td rowspan="2"></td></tr>
<tr><td>29</td><td>FR</td></tr>
</table>

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

31		RL			
32		RR			

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

34		FRh			
35		RLh			
36		RRh			
37	Hamasaki Sqaure (HS)	FL	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°
38		FR			
39		RL			
40		RR			

41	HS height layer at 0m	FL	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)

42		FR			
43		RL			
44		RR			

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

46		FR			
47		RL			
48		RR			
49	Side	SL	DPA 4011	Cardioid	*d*(SL-SR) = 5m; ∠(SR-centre line) = 135° (i.e., facing away from the source)
50		SR			
51		SLh	DPA 4018	Supercardioid	*d*(SLh-SRh) = 5m; ∠(SRh-SR) = 90°
52		SRh			
53	Voice of God	VOG	DPA 4018	Supercardioid	Facing directly upwards
54	Floor	FLf	AKG 414B-TLII	Cardioid	*d*(FCf-base) = 0.25m; *d*(FRf-FLf) = 1m; facing direfactly backwards
55		FRf			
56		FCf			
57	Dummy head	L	Neumann KU100	Omni	*d*(L-R) = 0.17m;
58		R			
59	ORTF	L	DPA 4011	Cardioid	*d*(L-R) = 0.17m; ∠(L-R) = 110°
60		R			
61	FOA	1	Sennheiser Ambeo	Raw (A-format)	See https://en-uk.sennheiser.com/microphone-3d-audio-ambeo-vr-mic
62		2			
63		3			
64		4			
65	Spherical (HOA)	32ch	mhAcoustics Eigenmike EM32	Raw (A-format)	See https://mhacoustics.com/sites/default/files/ReleaseNotes.pdf
66	Spots	Violin	Schoeps MK4	Cardioid	*d*(source to mic) = 0.7m; placed above the instrument, pointing towards the F hole.
67		Violin			
68		Viola	Schoeps CCM4		
69		Cello			
70		Piano L	Neumann KM184	Cardioid	d(L-R) = 0.5m; placed just outside the piano, pointing towards the hammers.
71		Piano R			