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

Multichannel 3D Microphone Arrays: A Review

HYUNKOOK LEE, *AES Fellow*
(h.lee@hud.ac.uk)

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

Along with the recent advance of multichannel 3D audio technologies, a number of new microphone techniques for 3D sound recording have been proposed over the years. To choose a technique that is most suitable for the intended goal of a recording, it is first necessary to understand the design principles, pros, and cons of different techniques. This paper first categorizes existing 3D microphone arrays according to their physical configurations, design philosophies, and purposes, followed by an overview of each array. Studies that have subjectively or objectively evaluated different microphone arrays are also reviewed. Different approaches in the configuration of upper microphone layer are discussed, aiming to provide theoretical and practical insights into how they can contribute to creating an immersive auditory experience. Finally, limitations of previous studies and future research topics in 3D sound recording are identified.

0 INTRODUCTION

0.1 Background

The last decade has seen a rapid growth in the research and development of the so-called "immersive" audio. The exact definition of the term immersion is still in debate among researchers. However, in the context of media and entertainment, immersive audio usually refers to three-dimensional (3D) audio that provides the listener with the sense of auditory height as well as the auditory width and depth, overcoming the limitations of the conventional stereo and surround sound systems. Over the years, several proprietary multichannel 3D audio formats have been developed, e.g., Dolby Atmos [1], Auro-3D [2], DTS:X [3], NHK 22.2 [4], etc., and the number of available content produced for such formats is gradually increasing. Furthermore, a set of 3D loudspeaker configurations for advanced sound systems have been recommended in ITU-R BS.2051-2 [5]. All of these systems commonly employ loudspeakers that are positively elevated from the listener's ear height, as well as those that surround the listener horizontally. Content for such systems can be created using one of the following approaches: channel-based audio (CBA)[1] , object-based audio (OBA)[1], scene-based audio (SBA)[1], or a combination of all of those. For example, CBA can be used to provide a "bed" of static auditory scene (e.g., ambience), while OBA is used for the flexible panning and level control of certain audio objects in reproduction. New technologies for transmitting or storing 3D audio content such as MPEG-H [6] and Dolby AC-4 [7] are also being widely deployed in broadcasting and entertainment sectors.

The new 3D audio formats necessitate new content production techniques. In the context of acoustic music and ambience recording, microphone technique is of paramount importance for rendering the spatial and tonal characteristics of auditory scene desired by the recording engineer and producer. While OBA deals with dry audio signals (e.g., synthetic or spot microphone signals) that can be flexibly mixed in post-production, a successful CBA recording using a microphone array in an acoustic environment requires careful considerations on various factors, such as the spacings and subtended angles among the microphones, microphone polar patterns, and the placement of the array.

[1] CBA refers to audio captured or rendered for a specific loudspeaker channel configuration. OBA relies on the metadata of each discrete audio object, which can be used for flexible panning and level control of each object for different loudspeaker configurations. SBA typically refers to audio captured or synthesized using the Ambisonics technology, which represents the physical sound field at a specific point in a space. See [90] for a more comprehensive overview about the above approaches.

A number of different 3D microphone array systems for CBA have been proposed by researchers and recording engineers since early 2000s, although most of them are from the last decade. Some of these systems have been designed based on certain scientific principles of auditory perception, whereas some others have been derived from best practices. In addition, Ambisonic microphone systems for SBA have become increasingly popular for 360° virtual/augmented reality applications over the recent years. Since it is unlikely that a single technique can serve all purposes in various different recording scenarios, it would be important for recording engineers and producers to understand the design principles of different techniques in order to be able to adopt a technique that is most optimal for their technical and artistic goals.

0.2 Aim and Scope

From the above background, the present paper aims to provide a comprehensive overview of existing multichannel 3D microphone arrays and discuss their physical and perceptual differences. The scope is limited to 3D microphone arrays proposed for the purpose of indoor music and outdoor ambience (e.g., urban soundscape and sport) recording and those that are compatible with 3D loudspeaker configurations recommended in ITU-R BS.2051-2 [5]. Hence, binaural microphone systems and microphone arrays specifically intended for sound source localization, beamforming, or sound field synthesis are not covered in this paper. Ambisonic microphone systems are covered in this paper since they are widely used for music/ambience recording as well as for localization and beamforming purposes. A number of different Ambisonic microphone systems are commercially available. However, since they share the same design concept and goal, the technical requirements of Ambisonics are discussed rather than reviewing each individual system.

To collect relevant publications that propose or evaluate multichannel 3D microphone arrays, the following database and conference proceedings were searched: Audio Engineering Society (AES) e-Library, Google Scholar, Proceedings of the International Conference on Spatial Audio, and Proceedings of Tonmeistertagung. Keywords used for the search include “3D microphone array,” “3D microphone technique,” “height microphone,” and “immersive recording.” From this, 18 novel multichannel 3D microphone techniques that are compatible with loudspeaker configurations in ITU-R BS.2051-2 [5] were identified, including four native Ambisonic techniques. It was also recognized that 3D microphone techniques devised by three professional recording engineers were introduced through workshop sessions at AES conferences or magazine articles but not published as papers. Hence, further information on those techniques were gathered through personal communications with the engineers as well as available online resources.

The rest of the paper is organized as follows. Sec. 1 categorises and overviews the existing 3D microphone arrays. Sec. 2 provides a brief review of studies that subjectively or objectively evaluated different systems. Sec. 3 discusses different types of upper microphone configurations identified from the review and their advantages and potential limitations from both technical and artistic standpoints. Additionally, future research required in the area of 3D sound recording is discussed.

1 OVERVIEW OF EXISTING TECHNIQUES

This section firstly discusses a channel and format labeling convention to be used throughout the paper. 3D microphone techniques that have been proposed for acoustic music and ambience recording are then categorized based on physical configuration, design philosophy and purpose, followed by the overview of each of them.

1.1 Channel and Format Labeling Convention

It is first necessary to clarify the loudspeaker channel and format naming convention to be used throughout this paper. Currently, there is an inconsistency in the labeling of individual channels and layers for different reproduction formats, which often causes confusion. For example, channels with the same purpose and similar loudspeaker positions are named differently, depending on the format, e.g., Left top front (Ltf), Front Height Left (FHL), or Top Front Left (TpFL). A group of positively elevated loudspeakers are interchangeably referred to as a “height,” “top,” or “upper” layer, while the ear-height speakers are described as a “main,” “middle,” or “base” layer. Furthermore, even an identical format is called differently, e.g., 9.1, 5.1.4, or 4+5+0. This is also true when referring to microphone arrays. Since most of the microphone arrays reviewed in this paper use a discrete microphone–loudspeaker routing, it is convenient to have a consistent labeling of microphone and corresponding loudspeaker channels. Arguably, the term “height” layer is most commonly used to describe the positively elevated loudspeakers in proprietary formats such as Dolby Atmos [1] and Auro-3D [2]. However, it might be an ambiguous name for formats that also employ negatively elevated loudspeakers, e.g., NHK 22.2 [4]. For the same reason, identification that extends the conventional 5.1 with the number of positively elevated loudspeakers (e.g., 5.1.4) would not be suitable for standardization.

For the purpose of consistency, this paper uses the labeling convention used in the ITU-R BS.2051-2 recommendation on advanced sound system for program production [5], which identifies a number of currently used and possible 3D loudspeaker formats in the form of “upper + middle + bottom” loudspeakers. For instance, 22.2 and 9.1 are described as 9+10+3 (System H) and 4+5+0 (System D), respectively. This is deemed to be a clearer way to identify different loudspeaker formats as the number of loudspeakers for each layer is indicated. Although this labeling convention does not indicate the number of low frequency effect (LFE) channels, it is adopted in this paper since microphone arrays used for acoustic recording typically do not employ an LFE channel.

This paper also uses the “set of parameters” (SP) labels from ITU-R BS.2051-2 for referring to loudspeaker chan-

Table 1. Categorization of 3D microphone arrays reviewed in this paper. Loudspeaker configurations based on ITU-R BS.2051-2 [5].

			Perceptually Motivated															Physical Motivated	
			Horizontally & Vertically Spaced (HVS)										Horizontally Spaced & Vertically Coincident (HSVC)					Horizontally & Vertically Coincident (HVC)	
Loudspeaker channel			Main							Amb.	Main/Amb.		Main				Amb.	Main	
Label	Azimuth (°)	Elevation (°)	OCT-3D 4+5+0	Bowles array 4+5+0	Williams Umbrella 4+4(8)+0	2L-Cube 4+5(7)+0	Spider Tree 4+5(7)+0	Twin Cube 4+4(5)+0	Double UFIX 4+5+0	Hamasaki Cube 4(5)+4+0	Hamasaki et al. 9+10+3	Howie et al. 9+10+3	PCMA-3D 4+5(7)+0	ORTF-3D 4+4+0	ESMA-3D 4+4(8)+0	aud3Dio 4+6+0	Lee Rec-3D 4+4+0	Native Ambisonic arrays	Tetra/ Spherical Ambisonic arrays
M+000	0	0	√	√	(√)	√	√	(√)	√		√	√	√		(√)				
M+030	+30	0	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√		
M–030	–30	0	√	√	√	√	√	√	√	√	√	√	√	√	√	√	√		
M+060	+45..+60	0									√	√							
M–060	–45..–60	0									√	√							
M+090	+90	0			(√)	(√)	(√)				√	√	(√)		(√)	√			
M–090	–90	0			(√)	(√)	(√)				√	√	(√)		(√)	√			
M+110	+100.. +120	0	√	√	√	√	√	√	√	√			√	√	√		√		
M–110	–100..–120	0	√	√	√	√	√	√	√	√			√	√	√		√		
M+135	+110.. +135	0									√	√				√			
M–135	–110..–135	0									√	√				√			
M+180	+180	0			(√)						√	√			(√)				
U+000	0	30..45			√						√	√						Decoding to flexible loudspeaker configurations	
U+030	30..45	30..45	√	√		√	√	√	√	√			√	√	√		√		
U–030	–30..–45	30..45	√	√		√	√	√	√	√			√	√	√		√		
U+045	45..60	30..45									√	√							
U–045	–45..–60	30..45									√	√							
U+090	90	30..45			√						√	√							
U–090	–90	30..45			√						√	√							
U+110	100..135	30..45	√	√		√	√	√	√	√			√	√	√		√		
U–110	–100..–135	30..45	√	√		√	√	√	√	√			√	√	√		√		
U+135	110..135	30..45									√	√							
U–135	–110..–135	30..45									√	√							
U+180	180	30..45			√						√	√							
T+000	–	90								(√)	√	√							
B+000	0	–15..–30									√	√							
B+045	45..60	–15..–30									√	√							
B–045	–45..–60	–15..–30									√	√							

nels. SP labels indicate the initial of the loudspeaker layer, followed by the azimuth angle of the loudspeaker (e.g., M+030, U-135, B+045, etc.). The ranges of allowed azimuth and elevation angles for the loudspeaker are provided in addition (see Table 1). Note that this convention follows the anticlockwise angular orientation, e.g., M+030 refers to a loudspeaker at 30° to the left from the listener position. For microphone arrays whose reproduction is based on one-to-one routings between microphones and loudspeakers, this paper will use the SP labels interchangeably for referring to both loudspeakers and microphones.

1.2 Categorization of Currently Available Techniques

This section categorizes existing 3D microphone techniques for music and ambience recording according to their physical configuration, purpose, and design motivation. Table 1 summarizes the 3D microphone arrays reviewed in this paper using the ITU-R BS.2051-2 SP labels.

1.2.1 *Physical Configuration*

In terms of physical configuration, current 3D microphone arrays broadly split into "horizontally & vertically spaced" (HVS), "horizontally spaced & vertically coincident" (HSVC), and "horizontally & vertically coincident" (HVC) arrays. The HVS arrays apply a certain spacing between all microphones, producing interchannel time differences (ICTDs) vertically as well as horizontally. It is well known that a wider horizontal microphone spacing would lead to a stronger spatial impression in reproduction due to a greater magnitude of interchannel decorrelation [8–10], which is one of the main reasons for the choice of a spaced microphone array. However, recent research suggests that vertical microphone spacing and vertical decorrelation would have a minimal or no effect on perceived spatial impression in 3D sound reproduction [11,12]. Based on this, some microphone arrays have been designed using the HSVC concept, where the middle and upper layers have no spacing while horizontally arranged microphones are spaced for enhancing spatial impression. This concept requires the use of directional microphones for the middle and upper layers for sufficient channel separation. Finally, arrays in the HVC group have no or minimal spacing between all microphones, thus relying on interchannel level differences (ICLDs) for directional source imaging. They are typically designed to utilize the Ambisonics technology [13,14]. Commercially available first-order Ambisonic (FOA) microphones systems consist of four cardioid capsules arranged in a tetrahedral layout, whereas higher-order Ambisonic (HOA) systems tend to have multiple small capsules installed on the surface of a small sphere, offering a higher spatial resolution than FOA. For FOA, it is possible to derive Ambisonic signals natively by configuring individual microphones, as will be introduced in Sec. 1.5.2.

1.2.2 *Design Philosophy*

The arrays can be also categorized into "perceptually motivated" and "physically motivated" arrays. The former arrays are typically designed to achieve certain characteristics in phantom image localization and spatial impression by manipulating the interchannel relationship between the microphone signals. That is, the main motivation is to provide the listener with a spatially and tonally pleasing aural experience while plausibly representing the sound field, rather than duplicating it in reproduction. Most HVS and HSVC arrays fall under this category. Each individual microphone signal of a perceptually motivated array is routed to its corresponding loudspeaker, and thus, the number of microphones in the array directly indicate the number of channels involved in the target reproduction format, e.g., nine microphones for 4+5+0. The arrays in this group are designed for specific loudspeaker configurations, following the CBA paradigm. Most of the perceptually motivated arrays tend to augment existing seven, five, or four-channel surround microphone arrays with additional microphones to feed the upper loudspeaker layer, thus forming a 4+7+0, 4+5+0, or 4+4+0 configuration.

On the other hand, physically motivated arrays are designed with the goal of sound field reconstruction. The aim of Ambisonic recording using a tetrahedral or spherical microphone array is to reproduce auditory localization cues in the physically most accurate way. The original definition of Ambisonics [13] requires three physical conditions to be met in the decoding process: (i) the directions of velocity and energy localization vectors [15] are the same at least up to around 4 kHz, (ii) the velocity vector magnitude (rV) is 1 below around 400 Hz, and (iii) Between 700 Hz and 4 kHz, the energy vector magnitude (rE) is maximized for as many directions of the 360° sound stage as possible. In essence, these conditions imply that interaural time differences (ITDs) below 400 Hz and interaural level differences (ILDs) above 700 Hz should agree in their direction up to at least 4 kHz [16].

1.2.3 *Purpose*

The microphone arrays are also divided into "main" and "ambience" arrays. A main array traditionally means a set of microphones that are arranged with certain spacings and subtended angles for both directional source imaging and creating spatial impression from the perspective of a specific location in the recording space. In the context of classical music recording, a main array would be typically placed at least about 2–4 m above the floor and slightly behind the conductor position, depending on the size of the ensemble and the desired spatial and tonal characteristics. In case of a 3D main array, the middle layer of the array would be responsible for source imaging in the front and ambient imaging in the rear. On the other hand, the upper layer would typically aim to capture ambience for enhancing perceived spatial impression, except when the sound source is physically elevated (e.g., choir on risers) or vertically large (e.g., pipe organ), in which case a vertical source imaging would be desired. Although some recording engineers rely heavily on a main array alone, others tend to add multiple "spot" microphones to complement the main array, but

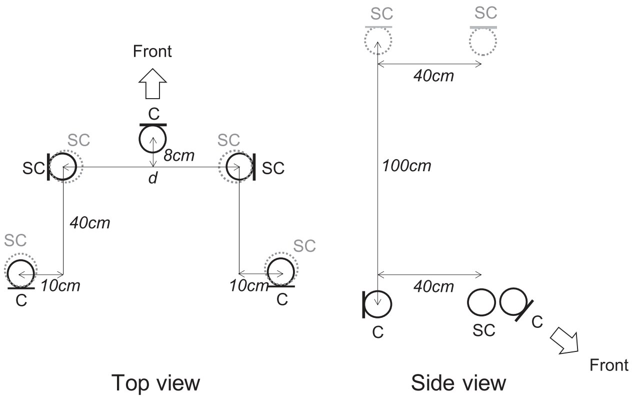

Fig. 1. Top and side views of OCT-3D. The solid black and dotted grey circles represent the middle and upper layer microphones, respectively. C = cardioid, SC = supercardioid.

spot microphone techniques are not within the scope of this paper.

On the other hand, ambience arrays are dedicated solely to capturing diffuse sounds (reflections and reverberation) rather than direct sounds, thus often being recommended to be placed beyond the critical distance of the recording venue [8]. An ambience array could be used in conjunction with the frontal middle layer microphones of a main array. A potential benefit of this approach is that there might be a larger headroom for raising the level of ambience without affecting the source image since the ambience array signals would contain little direct sound.

However, from a purist's point of view, it may be claimed that this approach effectively mixes the acoustic characteristics of two separate listening positions.

For a large reproduction format such as 9+10+3, some recording techniques involve both main and ambience arrays ("main/ambience" arrays). They typically use forward-facing main microphones exclusively for the directional imaging in a relatively close proximity to the sound sources and place multiple ambience microphones at considerable distances from them, thus requiring a large setup.

1.3 HVS Arrays

1.3.1 OCT-3D

The OCT-3D is a 4+5+0 array proposed by Theile and Wittek [17]. It uses the OCT-Surround (Optimized Cardioid Triangle Surround) 5-channel main microphone array [18] as the middle layer, augmented by four upward-facing supercardioid microphones at 1 m directly above it (Fig. 1). The main design aim of the front triplet of OCT-Surround is to achieve a stable frontal phantom imaging by minimizing 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 undesired crosstalk [18]. The suppression of ICXT with the front triplet is realized by using the sideward-facing supercardioids microphones. However, research suggests that the ICXT of a surround microphone array could contribute to an increase in perceived source width [19], which could be a positive perceptual effect depending on the type of sound source [20].

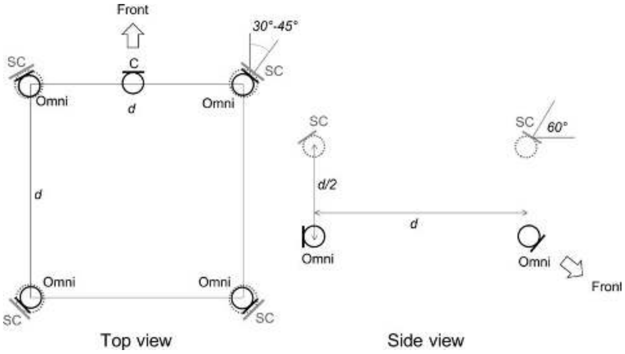

Fig. 2. Top and side views of Bowles Array. The solid black and dotted grey circles represent the middle and upper layer microphones, respectively. SC = supercardioid.

The choice of the supercardioid polar pattern for the upper layer is again for the reduction of ICXT. Theile and Wittek [17] asserts that it is not possible to achieve a stable vertical directional imaging between the middle and upper layers due to the limitation of vertical panning using ICTD or ICLD. This is supported by experimental results reported in several studies [21–23]. Theile and Wittek [17] suggest that the space between the middle and upper loudspeaker layers in reproduction should be filled with reflections and reverberation rather than the direct sound. Using upward-facing supercardioids for the upper layer would sufficiently suppress direct sounds, thus allowing mainly diffuse sound images to be rendered vertically.

1.3.2 Bowles Array

Bowles [24] also proposed to use supercardioid microphones for the upper layer to achieve a sufficient separation between the middle and upper layers. The middle layer of the Bowles 4+5+0 array (Fig. 2) uses four spaced omnidirectional (omni, hereon) microphones for M+030, M–030, M+110 and M–110 and one unidirectional microphone for M+000. He suggests that the spacing between each microphone can vary depending on the ensemble size. The upper

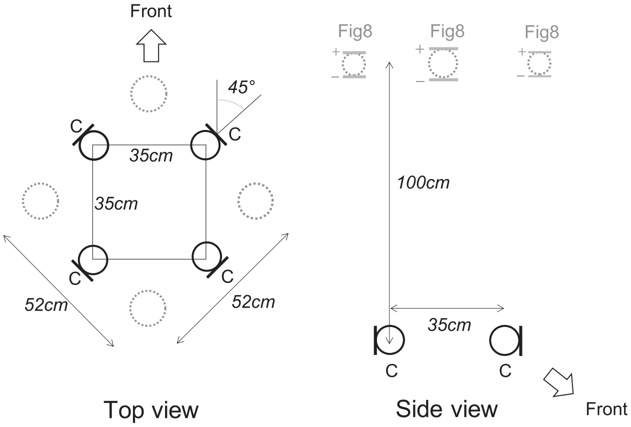

Fig. 3. Top and side views of Williams Umbrella. The solid black and dotted grey circles represent the middle and upper layer microphones, respectively. C = cardioid, Fig8 = figure-of-eight.

layer microphones are placed directly above the middle layer one, pointing around 60° upwards so that their null axis points directly below. They are also angled slightly outwards to achieve a better channel separation between one another. Bowles [24] recommends that the spacing between the middle and upper layers is one-third of the horizontal microphone spacing, claiming that too much vertical spacing might create a "hole in the middle" between the two layers. He also asserts that using omni microphones for the upper layer would cause an excessive low frequency presence and a poor localization, as well as a potential comb-filter effect. This is supported by some of the objective measurements conducted for various 3D microphone arrays [25].

1.3.3 Williams Umbrella

In contrast with OCT-3D and Bowles Array, Williams Umbrella (Fig. 3), proposed by Williams [26], aims to achieve a directional imaging on the diagonals between the middle and upper loudspeaker layers. The design of the array was motivated from an informal listening test using a two-channel stimulus with varying ICTD to ICLD ratios. During the test, Williams observed that it was possible to precisely localize the resulting phantom image along the vertical diagonal plane (e.g., between M–030 and U+030), while localization between directly vertical loudspeakers was poor. He further found that the diagonal localization precision was higher with a higher ICTD/ICLD ratio.

From these observations, he proposed the Umbrella array that has multiple isosceles triangle structures between the middle and upper layers, so that both ICTD and ICLD could be created between diagonally arranged microphones. As can be seen in Fig. 3, the middle layer consists of cardioid microphones arranged in a small square and pointing towards +45°, –45°, +135° and –135°, respectively. The spacing between each adjacent microphone is suggested to be 35cm [27]. The upper layer employs four vertically oriented figure-of-eight (fig8) microphones arranged in a 52 cm × 52 cm cross formation (0°, +90°, –90°, 180°), which are placed 1 m above the middle layer. Optionally, four "satellite" cardioid microphones can be added to the middle layer (0°, +90°, –90°, 180°), at 1.5 m from the base point of the array [27] for a 4+8+0 format compatibility. If these extra microphones are utilized, the upper layer has a square orientation instead of the cross in order to achieve the isosceles triangle structures.

Williams [26] states that the 1-m vertical layer spacing was considered to be adequate for sufficiently separating the middle and upper layers. However, its implication on the balance between the resulting ICTD and ICLD is not clear. The use of the upward-facing fig8 microphones for the upper layer is to achieve a directional imaging for sound sources located not only above the upper layer but also between the middle and upper layers. However, this causes an issue of a polarity inversion between the cardioid microphone of the middle layer and the back lobe of the fig8. For this reason, Williams [26] suggests that the directional imaging of a sound source located below about 26.5° elevation from the axis that is perpendicular to the line between the two microphones would be unstable.

1.3.4 2L-Cube

2L-Cube is a microphone array developed by Lindberg [28]. It employs nine omni microphones in a cube arrangement for a 4+5+0 reproduction (Fig. 4). The width and depth dimensions of the cube could vary from 0.4m to 1.2m depending on the size of the ensemble, while the height dimension is kept constant as 1m [29]. The front center microphone is placed slightly in front of the base point between the left and right microphones (the exact spacing is not specified.). In Lindberg's recording sessions, a large musical ensemble is usually arranged in a circular layout and 2L-Cube is placed at the center position to achieve a 360° source imaging. He also tends to adjust the distances of individual musicians from the microphone array in order to achieve an optimal level balance for each different musical work [29].

The choice of omni microphones for 2L-Cube is more for their tonality rather than the polar response itself [29]; an omni microphone would typically offer a more extended low-end in the frequency response compared to a unidirectional or bidirectional microphone. Furthermore, the exact vertical orientations of the upper layer microphones depend on the desired tonal characteristics [29]. He often utilizes acoustic pressure equalizers[2] to increase the directionalities of the microphones at high frequencies. This would produce some ICLDs vertically, which might be useful for avoiding the upward-shifting of source image in the vertical plane, which will be discussed in more detail in Sec. 3.1.

1.3.5 Spider Tree

As with Lindberg, Sawaguchi [30,31] arranges musicians in a circular formation around the main microphone array

[2] The acoustic pressure equalizers diffract sound waves near the microphone diaphragm, resulting in increases in upper-middle and high frequency energies as well as the directionality at high frequencies.

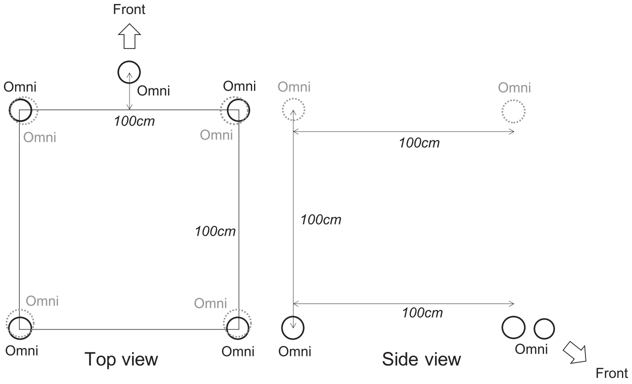

Fig. 4. Top and side views of 2L-Cube. The solid black and dotted grey circles represent the middle and upper layer microphones, respectively. All microphones are omnidirectional.

placed in the center of the ensemble. The motivation for this approach is to achieve a combination of realistic and creative surround sounds [31]. The array uses five omni microphones for the conventional 0+5+0 reproduction, with each microphone being 90–110 cm away from the array base point, and optional two outrigger omnis to feed M+090 and M–090 for the 0+7+0 format (Fig. 5). This is named "Spider Tree."

For a 3D extension of the tree, Sawaguchi has experimented with various upper microphone layer configurations for different musical styles. For example, four upward-facing cardioid microphones were placed in a square formation slightly outside a string sextet, with the aforementioned five-channel pentagon being in the center of the ensemble. This was to present early reflections rather than reverberation in the upper layer, which was deemed to be suitable for the music being recorded: *The Four Seasons* by Vivaldi and *The Art of Fugue* BWV-1080 by J. S. Bach. For Schubert's *Death and the Maiden*, on the other hand, four omni microphones were arranged in line with a wide spacing between each and they were placed in front of the ensemble. The inner two microphones were for U+030 and U–030, with the outer two for U+110 and U–110. It could be argued that, due to the close proximity to the sound sources and the large distances between the microphones, the inner microphone signals would likely create a strong directional imaging of the ensemble, while the outer signals might produce a large "hall in the middle." Nevertheless, Sawaguchi states that this microphone configuration provided a "dynamic musical representation" [31]. Another technique that he explored employs two pairs of 120°-coincident cardioid microphones, with each placed at the far left or right end of the concert hall. For each pair, one cardioid faced towards the stage, while the other was angled towards the audience area. The forward-facing microphones fed U+030 and U–030, whereas the backward-facing ones were routed to U+110 and U–110.

1.3.6 Twin Cube

Twin Cube, developed by Zielinsky [32], employs eight dual-output microphones arranged in a cubic layout with the dimensions of about 2 m, with each microphone facing directly forward (Fig. 6). A microphone for M+000 can be added optionally. The array was originally introduced as a recording technique for the nine-channel Auro-3D format, exclusively using the Sennheiser MKH800 Twin dual-output microphones (thus often called "Ambeo" Cube). A dual output microphone consists of two cardioid capsules that are arranged in a "back-to-back" fashion so that their two separate output signals can be combined or subtracted with a different mixing ratio to create different polar patterns flexibly. For example, combining them with an equal

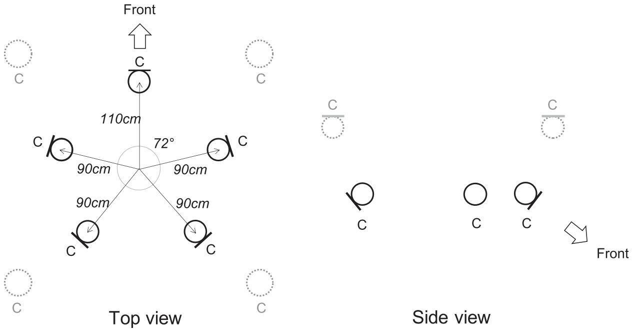

Fig. 5. Top and side views of Spider Tree. The solid black and dotted grey circles represent the middle and upper layer microphones, respectively. C = cardioid. Note that the upper microphone layer configuration shown in this illustration is one of various possible options described by Sawaguchi [30].

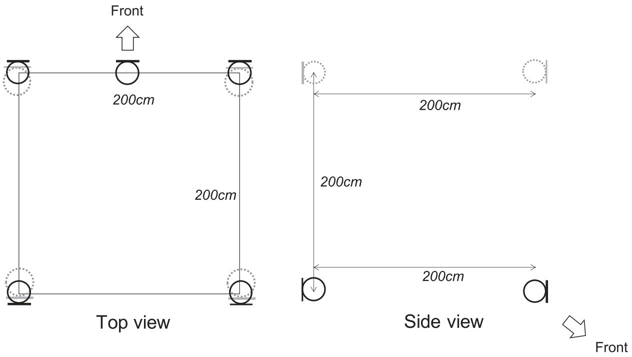

Fig. 6. Top and side views of Twin Cube. The solid black and dotted grey circles represent the middle and upper layer microphones, respectively. Each microphone has dual outputs allowing for creating a virtual microphone with a flexible polar pattern.

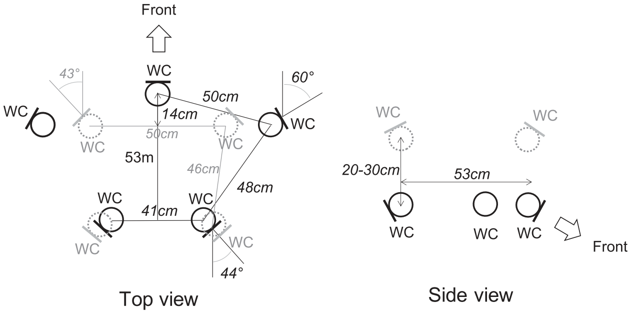

Fig. 7. Top and side views of Double-UFIX. The solid black and dotted grey circles represent the middle and upper layer microphones, respectively. All microphones are of the wide-cardioid polar pattern.

weighting would produce a virtual omni microphone, while subtracting one from the other would produce a virtual fig8. The same functionality can be achieved natively, by using a pair of individual cardioid microphones or a mid-side (MS) pair with an omni and a forward-facing fig8 at each corner of the cube.

Despite the flexibility of polar pattern control, the vertical orientation of the virtual microphone is restricted by the physical orientation, e.g., if the microphone is forward-facing as originally proposed, the resulting cardioid or fig8 microphones can only face forwards or backwards. The forward-facing cardioid or omni in the upper layer might produce a large amount of ICTX, potentially leading to an unstable phantom imaging in the vertical plane, as mentioned earlier. Its large physical dimensions, the array would be most suited for recording a large orchestra.

1.3.7 Double-UFIX

Double-UFIX is a 4+5+0 microphone array proposed by Camerer [33]. It consists of nine wide cardioid microphones that are configured as illustrated in Fig. 7. The main design motivation is to provide a good balance between listener envelopment (LEV) and phantom imaging accuracy for outdoor recording. The spacing and subtended angle for each microphone pair in the middle and upper layers are determined using the ICLD and ICTD trade-off functions provided in the MARRS (Microphone Array Recording and Reproduction Simulator) model [34], which allows for the calculation of stereophonic recording angle (SRA)[3] adaptive to loudspeaker base angle. The design aim is to match the SRA with the base angle of the loudspeaker pair that the microphones are routed to, following the critical linking concept by Williams and Le Du [35].

All of the nine microphones use the wide cardioid polar pattern since it provides a more extended low-frequency response compared with the cardioid. However, due to the relatively low directivity of the wide cardioid pattern, the accuracy of SRA matching might be influenced by ICXT. The resulting spacing between each adjacent microphone ranging from 41 cm and 50 cm would produce a full decorrelation of diffuse sound down to around 300 Hz according to a diffuse field coherence model provided in [36]. The vertical subtended angle between the middle and upper layer microphones is chosen to be 90° in order to produce a sufficient ICLD (e.g., 7 dB) based on [37]. The spacing between the middle and upper layers is recommended to be 25–30 cm since this would produce a full decorrelation of diffuse sound down to 500 Hz, below which the effect of decorrelation on vertical image spread is little [38].

1.3.8 Hamasaki Cube

Hamasaki and Van Baelen [39] vertically extended Hamasaki Square (HS) [8] for 3D ambience capture by adding an upper layer of four upward-facing supercardioids. HS is a popular technique used for capturing four-channel decorrelated ambient sounds for the conventional 0+5+0 reproduction. It consists of four sideward-facing fig8 microphones arranged in a square layout. Using both the front and rear channels for ambience recording and reproduction was found to produce a greater sense of LEV than using the rear ones only [40]. 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.

[3] 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.

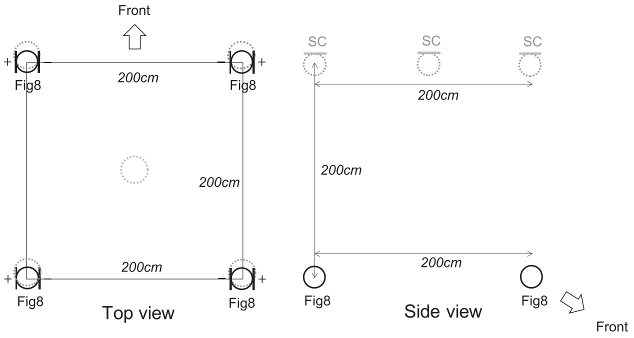

Fig. 8. Top and side views of Hamasaki Cube. The solid black and dotted grey circles represent the middle and upper layer microphones, respectively. SC = supercardioid, Fig8 = figure-of-eight.

The size of HS is recommended to be 2–3 m, based on subjective evaluation results and diffuse field coherence estimations based on [41], which suggests that a full interchannel decorrelation is achieved above 100 Hz with a 2-m microphone spacing. Hamasaki and Van Baelen propose using the same 2–3-m spacing between each adjacent supercardioid microphone in the added upper layer, as well as between the middle and upper layers. This forms a cubic layout (Fig. 8), referred to as Hamasaki Cube here. Additionally, an extra upward-facing supercardioid can be placed in the center of the upper square if an overhead loudspeaker (T+090, a.k.a. Voice of God [VOG]) is used as in the 9+10+3 loudspeaker format. A subjective evaluation that compared Hamasaki Cube against the original HS showed that the former was preferred to the latter overall [39].

1.3.9 Main/Ambience Array Approaches

In [42], Hamasaki et al. describe basic microphone arrangements used for recording an orchestra for the 9+10+3 format (i.e., NHK 22.2). A main array of five supercardioid microphones were arranged in a straight line and placed in front of the orchestra with equal intervals between each. For ambience capture, 13 fig8 microphones facing sideways were placed at various positions in the recording venue. The motivation behind the use of the widely spaced microphones was to achieve sufficient decorrelation between each microphone signal. However, information on the exact distances between the microphones and the routing between the microphones and loudspeakers is not provided in their paper.

Howie et al. [43] adopted and expanded Hamasaki et al.'s approach for experimental 9+10+3 recordings of a full orchestra made in a concert hall. As illustrated in Fig. 9, five front-facing supercardioid microphones were arranged in line behind the conductor position. Ambience was captured using eight sideward-facing fig8 microphones for the middle layer and eight upward-facing supercardioid microphones for the upper layer, which were arranged at widely spaced positions across the audience area of the concert hall. The upper layer ambience microphones were placed 3.67 m higher than the middle layer ones, which is consider-

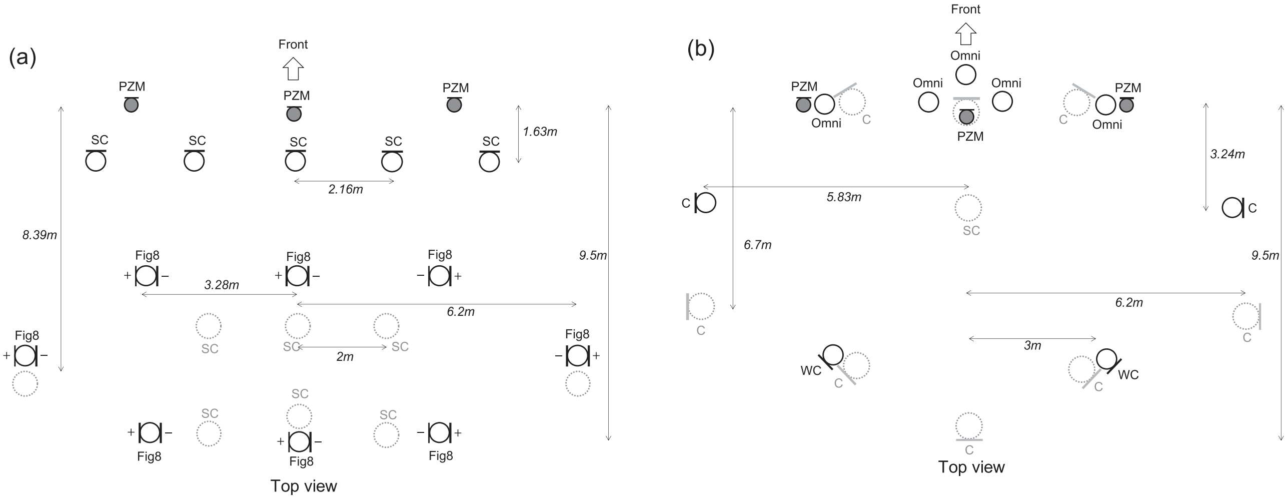

Fig. 9. Top and side views of Howie et al.'s 9+10+3 microphone arrangements used for recording a large orchestra. The solid black and dotted grey circles represent the middle and upper layer microphones, respectively. The filled grey circles represent bottom layer microphones. (a) is based on Hamasaki et al. [42]; middle layer height = 3 m, upper layer height = 6.7 m. (b) uses the Decca Tree as the main microphone array; middle layer height = 3 m, upper layer height = 5.5 m.

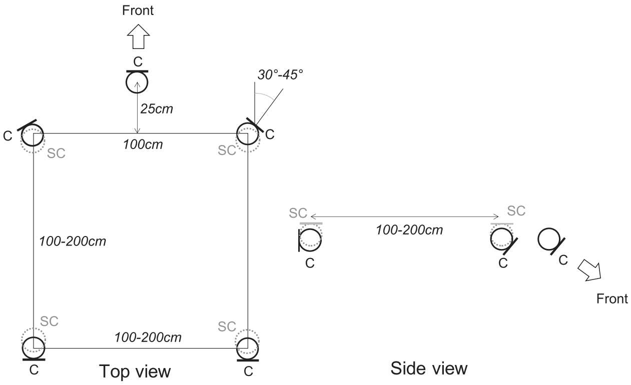

Fig. 10. Top and side views of PCMA-3D. The solid black and dotted grey circles represent the middle and upper layer microphones, respectively. C = cardioid, SC = supercardioid.

ably larger than the vertical spacing used in the other HVS arrays reviewed above. In addition, three widely spaced boundary microphones were placed on the stage floor to feed the bottom layer.

In the same recording session, Howie et al. also tested another main/ambience array technique using widely spaced microphones. The main array was a modified version of the popular "Decca Tree" microphone array using three omni microphones, with two extra omni outriggers. All other microphones apart from T+090 (supercardioid) were cardioids. The orientations of the microphones mirrored those of the loudspeakers, e.g., +60° azimuth and +45° elevation for U+045. Similar to Hamasaki et al.'s approach, the large setup with wide microphone spacings aimed for capturing decorrelated and diffuse indirect sounds.

1.4 HSVC Arrays

1.4.1 PCMA-3D

PCMA-3D [11,25] is a 4+5(7)+0 array, based on the PCMA (perspective control microphone array) design concepts proposed by the present author [44,45]. The original PCMA was devised for a flexible rendering of perceived distance and LEV in five-channel surround recording. Each point in the array employs forward-facing and backward-facing cardioid microphones arranged in a coincident fashion. By mixing the two microphone signals with a different ratio, a virtual microphone with different direction and polar pattern can be created, allowing the source-to-ambience ratio of each channel signal to be controlled flexibly.

This concept has been adapted to develop PCMA-3D based on two research findings: (i) the spacing between the middle and upper microphone layers (i.e., vertical interchannel correlation) would not have a significant effect on perceived spatial impression in 3D sound reproduction [11,12] and (ii) vertical interchannel time difference is an unstable cue for vertical phantom imaging [22]. These findings have also become the theoretical bases for the upper layer configurations of all other HSVC arrays, which are introduced in the following subsections.

A recommended configuration of the array is illustrated in Fig. 10. For the upper layer, four unidirectional microphones are coincidentally arranged with their corresponding middle layer microphones (e.g., M+030 and U+030). The vertical orientation of U+030 and U–030 are determined to ensure that the level of the direct sound captured by each of the microphones (i.e., vertical ICXT) is at least 9.5 dB lower than that of the corresponding middle layer microphone in order to prevent an unwanted upward shifting of the source image [37]. This can be easily achieved by facing cardioid or supercardioid microphones directly upwards. The null of supercardioid at around 127° is to be aligned with the on-axis of the sound source if a maximal suppression of vertical IXCT is desired. Alternatively, the cardioid microphones of the middle and upper layers can be arranged in a "back-to-back" fashion so that the upper cardioid faces away from the sound source for a maximal channel separation. These approaches allow the upper layer to mainly capture ambience arriving from the ceiling, while the middle 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. Another benefit of the vertically coincident configuration is that there would be little tonal coloration when performing a 3D-to-2D downmix. That is, little comb-filtering would occur when the height microphone signals are combined with their corresponding base microphone signals.

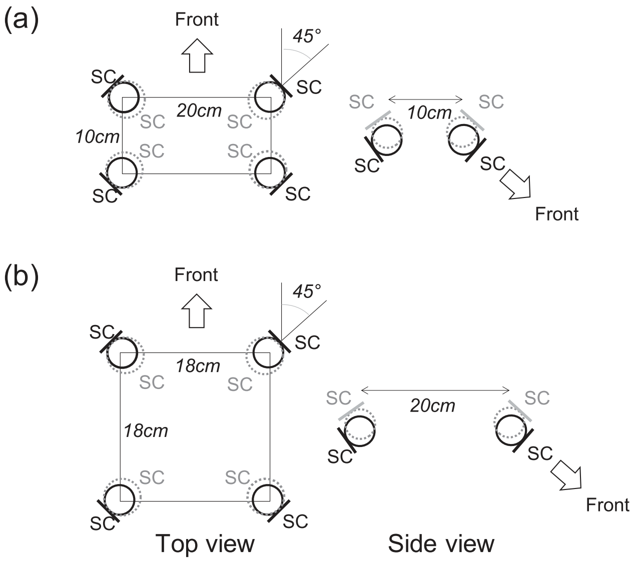

Fig. 11. Top and side views of ORTF-3D. The solid black and dotted grey circles represent the middle and upper layer microphones, respectively. SC = supercardioid. (a) is the outdoor set, and (b) is the indoor set.

1.4.2 ORTF-3D

ORTF-3D is a 4+4+0 array proposed by Wittek and Theile [46]. ORTF (Office de Radiodiffusion Télévision Française) is a popular two-channel near-coincident microphone technique, which uses two cardioid microphones with 17-cm spacing and 110° subtended angle. ORTF-3D by Schoeps uses a similar concept of narrowly spaced di-

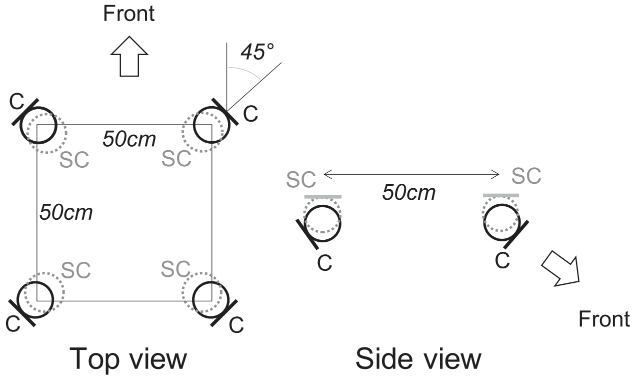

Fig. 12. Top and side views of ESMA-3D. The solid black and dotted grey circles represent the middle and upper layer microphones, respectively. C = cardioid, SC = supercardioid.

rectional microphones as illustrated in Fig. 11, employing four supercardioid microphones for each of the middle and upper layers that are arranged in a vertically coincident fashion based on [11]. This offers a compactness in the design, which makes it useful for outdoor location recording. The subtended angle between the middle and upper microphones is 90°, which provides a sufficient channel separation due to the high directivity of the chosen polar pattern. There exist two versions of ORTF-3D: outdoor set and indoor set. The indoor set has a square layout with 18-cm spacing and 90° subtended angle between each vertical microphone pair. On the other hand, the outdoor version has a rectangular shape, with 20-cm width and 10-cm depth.

1.4.3 ESMA-3D

Equal segment microphone array (ESMA) attempts to capture a continuous 360° sound field, which was originally proposed by Williams [47]. It is based on the "critical linking" design concept, which attempts to connect the SRA of each stereophonic segment of the microphone array without an overlap or missing gap. Williams suggested that an ESMA can be configured with a different number of cardioid microphones (e.g., four, six, and eight channels), provided that three conditions are met: (i) all pairs of adjacent microphones in the array must have an equal subtended angle, (ii) the subtended angle and the resulting SRA must be the same, and (iii) the loudspeaker array must have the same angular arrangement as the microphone array. For example, a four-channel quadraphonic ESMA must have the subtended angle of 90° between each adjacent microphone, which produces the SRA of 90°, and be reproduced over loudspeakers placed at ±45° and ±90° azimuth angles.

Williams originally proposed the microphone spacing for the quadraphonic ESMA to be 24 cm. This was based on his psychoacoustic model for SRA estimation, which was obtained using the conventional 60° loudspeaker base angle. However, it was found by the present author that, for the 90° base angle used in the quadraphonic setup, the conventional model did not provide an accurate imaging for the target SRA [48]. Based on the MARRS model [35], which applies a scale factor for the correction of SRA depending on the loudspeaker base angle, it was proposed to use the spacing of 50 cm for the quadraphonic ESMA (ESMA50). ESMA-3D [48] (Fig. 12) augments ESMA50 with four upward-facing cardioid or supercardioid microphones for the upper layer in a vertically coincident fashion, based on the same rationale used for PCMA-3D and ORTF-3D. In addition, an eight-channel octagonal ESMA-3D adds four extra microphones for the center, side, and back loudspeakers. This requires the microphone spacing between each adjacent microphone to be 53 cm.

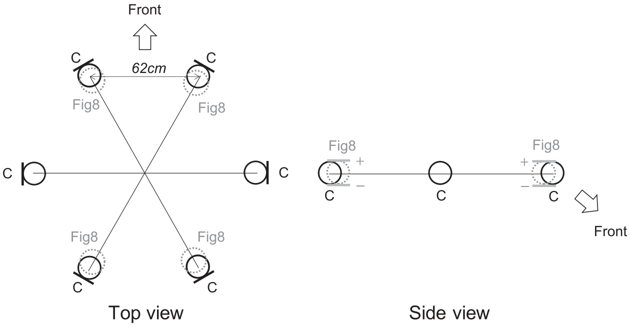

Fig. 13. Top and side views of au3Dio. The solid black and dotted grey circles represent the middle and upper layer microphones, respectively. C = cardioid, Fig8 = figure-of-eight.

1.4.4 au3Dio

au3Dio is a 4+6+0 array proposed by Vaida [49]. The middle layer consists of six cardioid microphones arranged in a hexagonal layout. The spacing between each adjacent microphone is recommended to be 62 cm to produce the SRA of 60° for each stereophonic segment based on a calculation using Sengpiel's psychoacoustic model [50]. The front and rear microphones are augmented with upward-facing fig8 microphones arranged in a vertically coincident fashion, as illustrated in Fig. 13. The fig8s can be directly routed to the upper layer loudspeakers, in which case the frontal cardioid microphones should face directly towards the sound sources to maximally suppress direct sounds in the upper layer signals. Alternatively, the cardioid and fig8 can be used as a vertical MS pair so that virtual microphone signals can be flexibly derived for the middle and upper layers. In this case, the microphone array needs to be raised above the height of the sound source.

1.4.5 Lee Rec-3D

Lee Rectangle(Rec)-3D [51] is an 4+4+0 HSVC ambience array with the dimensions of 1 m width and 0.5 m depth. The rationales for the rectangular layout as follow. Hamasaki Square [8] uses a square layout based on the implicit assumption that the effect of microphone spacing and interchannel decorrelation on LEV is equally important for both width and depth dimensions of the array. However, in the present author's recent experiment comparing different combinations of widths and depths (0.5 m, 1 m, and 2 m for each) of a four-channel ambience array [51], no significant difference was found among the three spacings in the depth dimension on either perceived depth, width or LEV, which suggests that the array depth could be as small as 0.5 m. For the width dimension, the 1-m and 2-m array widths did not have a significant difference on any of the three attributes, whereas the 0.5-m width produced significantly less perceived width and LEV. From this, the maximum necessary width and depth were determined to be 1 m and 0.5 m.

Based on the above and inspired by the approaches described in [8], three options are proposed for the choice of polar pattern for the middle layer (Fig. 14), depending on the desired spatial characteristics: (i) four backward-facing cardioids, (ii) four sideward-facing fig8s, and (iii) two sideward-facing fig8s and two backward-facing cardioids for the front and rear channels, respectively. The backward-facing cardioids would be useful for capturing more of reverberant sounds arriving from the back of a concert hall, which may sound warmer and provide a greater environmental depth than those from the side. On the other hand, sideward-facing fig8s would tend to capture stronger early lateral reflections, which might be beneficial for enhancing the perception of source and environmental width. The upper layer employs upward-facing supercardioids, which are placed coincidentally with the main layer.

1.5 HVC Arrays

1.5.1 Tetrahedral and Spherical Microphone Systems

Today, a number of Ambisonic microphone systems are commercially available (Table 2). In loudspeaker reproduction, FOA recordings typically suffer from a narrow optimal listening area (sweet spot). Ideally, the capsules used in an Ambisonic microphone system need to be arranged in a perfectly coincident fashion to encode spatial information accurately for the entire frequency spectrum. However, most commercial systems do not achieve this due to constraints in physical design. The intercapsule spacing leads to the so-called "spatial aliasing" at high frequencies [52]. As shown by Kurz et al. [53], different FOA systems suffer from different degrees of spatial aliasing, and this could lead to significant differences in terms of overall sound quality and preference as well as localization accuracy and LEV in reproduction. Bates et al. [54] also found that different Ambisonic systems vary considerably in localization accuracy and tonal quality. In addition, the size of the sweet spot could be significantly increased with HOA, as found in [55].

Table 2. List of commercially available Ambisonic microphone systems.

Maximum Ambisonic order	Commercially available Ambisonic systems
1st	Brahma Field 4 Core Sound TetraMic MiniDSP ambiMIK-1 Oktava MK-4012 Reynolds A-type 4 Rode NT-SF1 Soundfield SPS200 Soundfield ST450 MK2 Soundfield DSF-B MK2 Soundfield DSF-2 MK2 Sennheiser Ambeo VR Zoom H3-VR
2nd	Brahma Field 8 Core Sound OctoMic Reynolds A-type 8 Voyage Audio Spatial Mic
3rd	Zylia ZM-1
4th	mhAcoustics Eigenmike EM32
6th	mhAcoustics Eigenmike EM64

The raw signals from an Ambisonic microphone system is traditionally called the "A-format." They need to be converted into the "B-format" (i.e., spherical harmonics), which are then decoded according to loudspeakers. Some practical software tools accompanying commercial Ambisonic microphones often allow the user to create virtual microphones of different polar patterns for different output channels and flexibly steer them towards different directions to control the spatial characteristics of the resulting phantom image. Although this is a convenient and creative way of using B-format signals, it is important to note that this "beamforming" approach should not be confused with the original Ambisonic decoding. As already mentioned in Sec. 1.2.2, the original aim of Ambisonic decoding is to accurately reconstruct the ITD and ILD cues at the listening position. It is recommended in [13] and [16] that "velocity" or "basic" decoding, which is optimized for ITD matching, is used for frequencies below 400 Hz, whereas "energy" or "Max-r_E" decoding is used for the higher frequencies to achieve ILD matching. For an optimal Ambisonic decoding, loudspeakers must be configured in regular polygons or a polyhedral layout [16]. For irregular loudspeaker layouts, which are commonly found in commercial formats as well as the formats specified in ITU-R BS.2051-2 [5], alternative decoding methods such as energy-preserving Ambisonic decoding (EPAD) [56] and all-round Ambisonic decoding (ALLRAD) [57] have been proposed. Detailed reviews and

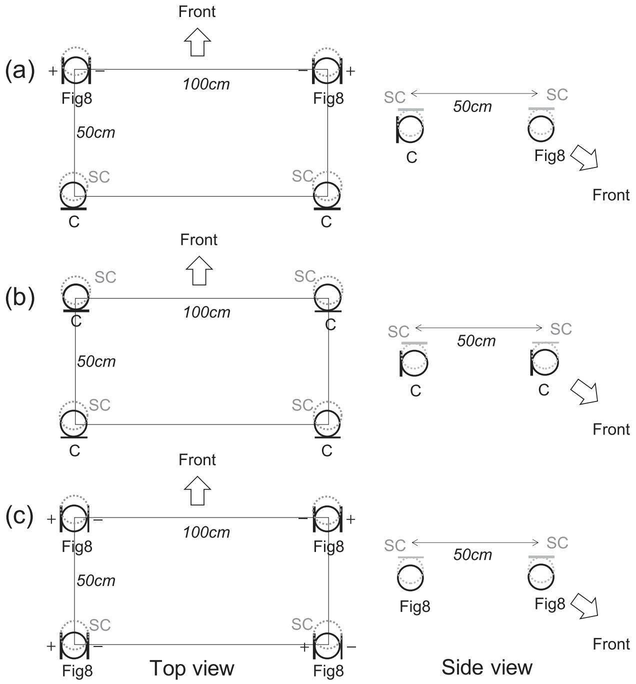

Fig. 14. Top and side views of the three versions of Lee Rec-3D. The solid black and dotted grey circles represent the middle and upper layer microphones, respectively. C = Cardioid, SC = supercardioid, Fig8 = figure-of-eight.

discussions of different Ambisonic decoding methods can be found in [14] and [16].

1.5.2 Native FOA Microphone Arrays

It is possible to create FOA B-format signals natively by using individual cardioid and/or fig8 microphones. Double MS is arguably the most popular native 2D FOA array. It consists of forward and backward-facing cardioid microphones arranged in a back-to-back coincident fashion (180° subtended angle) with a sideward-facing fig8 placed right above or below them. Summing the two cardioids produces a virtual omni microphone (i.e., W of B-format), whereas subtracting the backward-facing cardioid from the forward-facing one creates a virtual forward-facing fig8 (i.e., X). The sideward-facing fig8 serves as the Y component. Geluso [58] proposed to add an additional upward-facing fig8 microphone (i.e., Z) to the Double MS system for 3D recording, calling it Double MS+Z. A similar native 3D FOA system using three fig8s arranged for X, Y, and Z and an omni for W is referred to as Triple-MS by Zotter and Frank [14]. Another possible way for native B-format recording is to arrange four cardioids directly in a tetrahedral layout, employing a similar A-to-B conversion process similar to commercial tetrahedral arrays [14].

Additionally, Zhang and Geluso [59] proposed a microphone technique named three dual coincident capsules (3DCC), which can produce up to 18 output signals as well as native B-format signals, using three dual-capsule microphones arranged in a coincident fashion. By applying different weightings for different capsules in the summing matrix, different polar patterns between omni and fig8 are created flexibly for different output channels, aiming for a maximal channel separation between adjacent virtual microphones.

The main advantage of the native FOA approaches is that recording engineers can choose their favorite microphones rather than using the fixed capsules of a commercial FOA system. Furthermore, with a native array, it may be possible to arrange the microphones with a smaller intercapsule gap than some commercial FOA systems, which would potentially help reduce the spatial aliasing at high frequencies.

2 COMPARISONS AMONG 3D MICROPHONE ARRAYS

Some of the 3D microphone arrays described above have been compared against each other in subjective listening tests or objective measurements. Since the playback systems, types of sound source, and recording environment (as well as listening test questions) used in these studies vary largely, it is not possible to derive general conclusions on the differences between all of the individual arrays included in those studies. However, from the reviews of the available studies, it is possible to gain broad insights on perceptible differences between the three different types of physical configuration: HSV, HSVC, and HVC.

Scuda et al. [60] conducted subjective comparisons between PCMA-3D and Williams Umbrella in a 4+5+0 reproduction in terms of four perceptual attributes: the azimuth and elevation of phantom image, apparent source width (ASW), and room-related spaciousness. The stimuli were the recordings of castanets, cello, and speech made in a relatively dry large room. They found that PCMA-3D was generally more effective for shifting the image position in both azimuth and elevation planes. This might be related to the vertical polarity inversion issue of the Umbrella array that was described in Sec. 1.3.3. For ASW and spaciousness, there was no statistically significant difference found between the two arrays. This result supports the finding from Lee and Gribben's experiment [11]: the spacing between vertically oriented microphones did not influence overall spatial impression even though different vertical microphone layer spacings (0 m, 0.5 m, 1 m, and 1.5 m) produced considerable differences in the amount of vertical interchannel decorrelation.

Riitano and Victoria [61] subjectively compared three microphone arrays: ORTF-3D, Double MS-Z, and a 4+6+0 HVS array consisting of six cardioids for the middle layer and four supercardioids for the upper layer. They used four different types of recordings made with static and moving musicians around the arrays and tested them for locatedness (i.e., ease of localization), ASW, and LEV. There was no vertical aspect evaluated in their study. The results generally showed that ORTF-3D and the 4+6+0 array generally produced higher ratings than Double MS-Z in localization quality and LEV. On the other hand, there was no consistent trend found for ASW. Additionally, they reported that Double MS-Z suffered from phasiness. However, in this study, a different loudspeaker format was used for the reproduction of each array. Furthermore, since vertical localization or spatial impression was not evaluated in their study, it is not clear how the upper microphone signals contributed to the perceived results.

Double MS-Z was also compared against Twin Cube in a 4+5+0 reproduction by Ryaboy [62]. Recordings of a five-piece band consisting of violin, clarinet, voice, accordion and upright bass as well as a solo marching band style drum were used as stimuli for a listening test. Although statistical significance is unclear from the report, Twin Cube was generally found to have higher ratings than Double MS-Z in perceived ensemble width, perceived room size, and the sense of height, whereas the latter had smaller degrees of localization blur.

Perceived differences between HVC and HVS arrays were also reported by Howie et al. [43], who compared an HOA microphone system (mhAcoustics Eigenmike EM32) against the two types of 9+10+3 arrays described in Sec. 1.3.9 with orchestral recordings. The 3rd-order B-format signals from the HOA system were decoded to the 9+10+3 loudspeaker system (i.e., NHK 22.2) using a dual-band ALLRAD [57]. Their results suggested that the HOA recording had significantly lower ratings than the two large arrays in all of the attributes tested: clarity, environmental envelopment, naturalness, scene depth, and quality of orchestral image. A similar result was obtained from another experiment by Howie et al. [63], which compared Double MS-Z, OCT-3D, and two widely spaced 4+5+0 arrays: Double MS-Z was rated significantly lower than the other arrays in terms of ASW, LEV, and naturalness.

In their subjective study that compared solo violin recordings made using different microphone arrays, Kamekawa and Marui [64] found that an FOA system (Sennheiser Ambeo VR) was rated significantly lower than a 9+10+3 array based on [42] in terms of preference and presence (i.e., sense of being there) at off-center listening positions, whereas there was no significant difference between them at the central position (i.e., sweet spot). This result seems to be contradictory to the negative result obtained for the more advanced HOA system in the aforementioned study by Howie et al. [43]. Although the tested attributes were not exactly the same, the naturalness and environmental envelopment attributes tested in [43] seem to be related to the presence attribute used in [64]. The difference between the results from the two studies may be due to different experimental conditions used in those studies (e.g., sound source type, Ambisonic decoding method used, and acoustics of the recording venue and the listening room) or may be related to some inherent limitations of the HOA system. However, a potential influence of cultural differences between the Canadian [43] and Japanese [64] subject groups on the results would be worthwhile to consider. Although their study was not about microphone arrays, Kim et al. [65] reported the existence of cultural differences between Japanese and North American (Canada and US) subject groups in the comparison of different upper loudspeaker layer configurations. Furthermore, it was found in [66] that listeners' consistencies in the evaluation of 3D sound recordings depended on their audio production and musical training experiences.

The studies discussed above compared different microphone systems but did not exclusively consider the influence of the upper microphone layer on different perceived attributes. Some insights into this are provided from the objective comparisons between various 3D microphone arrays conducted by Lee and Johnson [25]. From multichannel room impulse responses captured using eight different arrays, ear-input signals for a 4+5+0 loudspeaker reproduction were derived, and various parameters were calculated. Two main differences in the ear signals occurred between different arrays as results of adding the upper layer signals. Firstly, PCMA-3D, 4th-order, and 1st-order Ambisonic systems (both decoded using ALLRAD [57]) produced considerably less spectral distortions in the ear-input signals than the HVS arrays, which exhibited predominant comb-filter patterns in the frequency spectrum. Secondly, the upper layers with omni microphones caused greater magnitudes of ITD and ILD fluctuations over time than PCMA-3D and the two Ambisonic systems, as well as greater levels of vertical ICTX. These results seem to indicate that not only vertically oriented spatial differences but also tonal and horizontally oriented spatial differences should be considered when discussing possible differences between 3D microphone arrays.

It is worth pointing out that, even though it is possible to decode Ambisonic recording to an irregular loudspeaker array using a method such as ALLRAD [57], an optimal result for Ambisonic decoding is achieved when a regular loudspeaker array is used [16]. However, all of the aforementioned studies that compared different microphone arrays used ITU-R–recommended irregular loudspeaker arrays for which the HVS or HSVC arrays tested have been designed specifically. Therefore, it does not seem entirely fair to judge the perceived quality of an Ambisonic microphone system against a perceptually motivated one based on the results reported in the studies.

To the author's best knowledge, a study conducted by Millns and Lee [67] is the only study that compared Ambisonic and spaced microphone arrays using a regular loudspeaker array. They compared concert hall recordings made using a four-channel ESMA (i.e., only the middle layer from ESMA-3D, illustrated in Fig. 12) and a Sennheiser Ambeo VR FOA microphone in terms of four perceived spatial attributes: source/ensemble width, source/ensemble distance, environmental width, and environmental depth. In their listening tests, the recordings were presented over headphones, where a virtual quadraphonic loudspeaker array (i.e., loudspeakers at ±45° and ±135°) were binaurally synthesized. Although this study did not use a 3D reproduction over physical loudspeakers, the test results show interesting differences between the two systems. Overall, ESMA was found to produce a greater source/ensemble spread, source/ensemble distance, and environmental spread than FOA. However, FOA was perceived to provide a similar or greater magnitude of environmental depth compared with ESMA, depending on the sound source position or the physical layout of the ensemble. Although this kind of comparison also needs to be conducted over physical loudspeakers to be able to draw a clearer conclusion, the results seem to suggest that different types of arrays produce different kinds of merits in terms of perceived spatial characteristics.

3 DISCUSSIONS

From the overview of the various different 3D microphone techniques, three main areas for discussions with respect to the use of the upper microphone layer have been identified: (i) the microphone polar pattern and orientation in the upper layer, (ii) the vertical spacing between the microphone layers, and (iii) realism versus artistry in the upper layer configuration. From the discussions of these points, it is aimed to provide both theoretical and practical insights into how the existing techniques could be adopted or adapted to achieve the goal of immersive auditory experience. Furthermore, limitations of the current research on 3D sound recording and reproduction are identified, and required future studies are proposed. Note that the present discussions focus on the configuration of the upper microphone layer. Detailed reviews and discussions on conventional surround microphone techniques for the middle layer can be found in [40] and [68]

3.1 Polar Pattern and Orientation of Upper Layer Microphones

The HVS and HSVC arrays reviewed tend to further split into two groups based on the configuration of the upper microphone layer: (i) those that use unidirectional microphones facing directly upwards or away from the sound source (e.g., [11,24,39,46,69]) and (ii) those that use omni or forward-facing unidirectional microphones (e.g., [28,30,32]). Both groups commonly use the same polar pattern for all of the upper layer microphones. The implicit assumption for the first group is that the upper layer is primarily for capturing realistic environmental sounds rather than for vertical phantom source imaging. Most of the arrays in this group aim for sufficient channel separations between the main and upper layers. This approach seems to be logical for recording sound sources located at a lower position than the microphone array, which would likely be the most common setup scenario for recording a small to medium classical ensemble. The minimum amount of vertical ICLD required for the phantom image to be localized around the height of the middle loudspeaker layer is around 7 dB for a vertically spaced microphone pair and around 9.5 dB for a vertically coincident pair [37]. This requirement would be easily met with upward-facing supercardioid or cardioid microphones when the array is raised to a sufficiently high position compared to the sound sources. This approach would also allow the sound engineer to raise the level of the upper channel signals for boosting the spatial impression without affecting the perceived position of the source images.

In case of using omni or forward-facing cardioid microphones for the upper layer, on the other hand, the upper layer would capture a considerable amount of direct sounds (i.e., vertical ICXT) as well as ambient sounds. This can lead the perceived source image to be shifted towards the upper loudspeaker layer [22,37]. Moreover, some tonal coloration may also be audible due to comb-filtering resulting from vertical ICTD [25]. Therefore, raising the level of the upper layer microphones higher than that of the main layer ones may worsen these potential problems and also increase the perceived loudness. Whether this issue matters or not in practice seems to be a debating topic [32,70] and will be further discussed in Sec. 3.3.

From a different perspective, the direct sounds captured by the upper layer microphones might make the arrays in the second group more effective for creating a vivid sense of height for physically elevated musical sources (e.g., instruments or choral singers on risers in a large-scale orchestra or a large pipe organ). It is worth noting, however, that even for a sound source elevated highly, it may be difficult to stably locate an image fully around the physical height of the upper loudspeaker layer since ICTD is not an effective cue for vertical image localization; the precedence effect would not fully operate in the vertical plane [22].

The extended low frequency response of omni microphones is often the main reason why some engineers prefer them to directional microphones [28]. For the same reason, in conventional stereo or surround recording, a unidirectional microphone is sometimes accompanied by an omni that is placed right above or below it (the so-called "Straus Paket" technique). Theile [18] proposed this approach as an option for the front left and right channels of OCT, such that the supercardioid and omni signals are combined together with high-pass and low-pass filtering applied at 100 Hz, respectively. Although this approach might be useful also for other arrays using directional microphones when recording sound sources with an extended low-end energy (e.g., full-scale orchestra or pipe organ), the usefulness of omni for the upper channel microphones is questioned. Due to the pitch-height effect [71,72], the low-frequency components of upper channel signals would likely be localized at the ear height or below. Furthermore, if both the middle and upper layers use omnis exclusively, the overall energy of low-frequency sounds and the amount of reverberation captured may become excessive as Bowles [24] points out.

Another potential advantage of using an omni upper layer is that a large vertical image spread (VIS) of the phantom source could be perceived [73]. The VIS might be a desired characteristic depending on the artistic intention of the recording. If a large VIS is desired and yet the upward-shifting of the image's focal point should be reduced, this could be achieved using several practical methods. For example, acoustic pressure equalizers could be adopted for the omni microphones to increase the directionality at high frequencies as used by Lindberg [28] and King et al. [70], depending on the availability for the microphone model used. Nipkow [74] uses a thick towel behind and an omni ambience microphone for the back channels to avoid excessive direct sound. This approach could also be adopted for an omni upper layer depending on the practicality in a given recording situation. Alternatively, omni upper layer signals could be equalized to reduce the energies of mid-high frequencies as a post-processing technique. This is based on the research findings that a large VIS is created mainly at frequencies below around 1 kHz [38], while the localization issue due to vertical ICXT is mainly due to frequencies above 3 kHz [73].

As found in a study by Howie et al. [75], listeners' preference on the upper layer polar pattern can be highly subjective. Therefore, rather than simply adopting a single technique in all recording situations, the polar patterns of upper layer microphones should be chosen carefully based on the producer or sound engineer's intended technical and artistic goals, which is further discussed in Sec. 3.3.

3.2 Vertical Microphone Layer Spacing

From the overview of the existing 3D microphone techniques, it was found that the main motivations for using a vertical spacing between the middle and upper layers applied in HVS arrays are as follows: (i) vertical imaging of diffuse sounds [24,69], (ii) vertical source imaging using ICTD [26], and (iii) interchannel decorrelation between the two layers [28,32]. It is well known that decorrelation is effective for horizontal image spread (HIS) [10] and the enhancement of LEV [8]. Most of the HVS arrays seem to follow this notion for vertical microphone spacing. However, the HSVC arrays use a vertically coincident arrangement between the two layers based on the research findings on the effects of vertical microphone spacing [11], interchannel decorrelation [12], and interchannel time difference [22] as mentioned in Sec. 1.4.1. The HSVC concept allows the arrays to be made physically more compact, thus providing practical advantages in outdoor soundscape or live concert recording situations. Another important advantage of the concept is in 3D-to-2D downmix. When the upper layer signals are directly mixed with their corresponding microphones in the middle layer for 2D reproduction (e.g., U+030 and M+030, U+110 and M+110), there would be little comb-filtering at the ear positions due to the coincident nature (i.e., no time delay between direct sound components). This would eventually cause less distortion in the ear-input frequency spectrum of the middle layer signals compared with HVS arrays, especially the ones using omni upper microphones [25]. Although the degree of decorrelation between vertically coincident microphones in the HSVC arrays would be smaller than that in the HVS arrays, a sufficient amount of decorrelation can still be achieved between vertically diagonal pairs (e.g., M+030 and U–030) [25].

On the other hand, an HVS array would have advantages over an HSVC one in terms of creativity and flexibility, especially for a large reproduction system (e.g., 9+10+3). The polar patterns and placements of the microphones could be chosen more flexibly due to the spacing. Although the absolute magnitude of VIS increase due to interchannel decorrelation is minimal [12], the effect can still be significant at frequencies above around 1 kHz, depending on the type of sound source [38]. Based on Kuster's diffuse field coherence model [76], the minimum required vertical microphone spacing to achieve a full decorrelation down to 1 kHz is 0.15 m for an omni pair and 0.1 m for a cardioid pair. Increasing the spacing further to 0.3 m and 0.2 m, respectively, results in a full decorrelation down to 500 Hz. This implies that the vertical spacing does not necessarily have to be as large as 1 m in order to benefit from decorrelation. A larger spacing might rather be useful for controlling tonal characteristics. When recording a large pipe organ in a church, for example, the spectral balance of the upper microphone signals can greatly vary depending on the vertical positions of the microphones. In such a recording scenario, it would be desired to create a source image in the upper layer as mentioned above. The upper layer height can be used as a creative tool to produce a desired tonal quality of the source image.

3.3 Realism Versus Artistry in 3D Microphone Techniques

While Ambisonic microphone systems attempt to capture and reproduce sound field as realistically as possible, motivations behind the configuration and placement of HVS and HSVC arrays tend to lie between realism and artistry. For example, those techniques that use a fixed configuration of upward-facing directional microphones seem to aim for a more realistic representation of the recording venue's acoustic characteristics, regardless of musical style. On the other hand, Sawaguchi [31] experimented with different microphone polar patterns and placements for the upper layer depending on the intended musical expression as introduced in Sec. 1.3.6.

Both realism and artistry are important aspects of sound recording, and they might complement each other in creating an immersive auditory experience, which is often described as the ultimate goal of 3D audio [77]. A number of studies (e.g., [78–80]) suggest that a technology-mediated immersive experience necessitates both the sense of being there (i.e., physical presence [81]) and the user's involvement in the narrative of the content. The microphone technique used for 3D recording can play an important role in inducing physical presence. However, it can be suggested based on [82,83] that the sound field of the recording venue would not necessarily need to be duplicated exactly, but a perceptually plausible representation of the auditory environment would be sufficient for providing the sense of being there. This leads to a discussion on the choice of microphone technique between physically and perceptually motivated arrays. Although Ambisonic microphone systems attempt to reproduce the sound field as authentically as possible, perceptually motivated HVS or HSVC arrays tend to produce higher subjective ratings on presence and naturalness as found by studies reviewed in Sec. 2. This is possibly related to the inherent limitation of the current Ambisonic technologies, such as tonal coloration [53] and narrow sweet spot [55]. From another viewpoint, however, it can be argued that listeners might feel more present with a sound that agrees better with their internal references about what realistic or natural sound should be. This might be the case especially because the listeners usually have no reference of the original sound field to compare the recorded sound field against. This is where subjective factors such as the listener's previous experience, expectation level, or cultural background would come into play in immersive experience.

The above discussion also poses a question of how the upper layer of a 3D microphone array should be configured to produce a strong sensation of being there. Defining and quantifying auditory factors that determine the level of physical presence in 3D reproduction is a complex topic. However, it seems logical to consider that a certain level of plausible representation of acoustic cues should be achieved in order to satisfy the listener's expectation about how the auditory space should sound. Revisiting the discussion about the debate on the use of omni microphones for the upper layer, an extreme upward-shifting of a solo violin image due to a strong vertical ICTX, for example, might be a potential factor that hinders the sensation of physical presence because in a concert hall the musician would not normally be positioned above the listener unless the listener sits in the very front row that is at a lower level than the stage. In this sense, it seems to be more reasonable to use upward-facing unidirectional microphones to capture reflections and reverberation from the ceiling.

From the viewpoint of listener involvement, however, various creative microphone techniques for the upper layer could be devised. In a concert hall recording, for example, cardioid microphones facing towards the audience area, which tends to absorb more high frequencies than the relatively untreated ceiling, may capture reflections and reverberation with softer characteristics than supercardioids facing directly upwards. Such differences in tonal color may elicit different emotional responses [84]. Pätynen and Lokki [85] reported that concert halls with stronger lateral sounds tend to increase the listener's emotional response for orchestral music. Based on this, orienting directional microphones of the upper layer towards the side walls to primarily capture the lateral sounds of the recording venue may help increase the listener's emotional responses and ultimately help him or her feel more deeply involved in the listening activity. Furthermore, the extended VIS and HIS resulting from omni upper layer microphones may be useful for abstract musical expressions, which would be beneficial for certain types of music (e.g., electroacoustic/acousmatic music). Ryan [86] states that an intense spatial immersion (envelopment) in virtual reality could develop an intimate relation to the narrative of the content as well as a sense of being there. Based on this, it could be suggested that a microphone array that can produce a stronger LEV might induce a higher degree of involvement even if it is exaggerated or different compared to real life experience.

In addition, a creative microphone placement may also help enhance the sensation of social presence [81] as well as involvement. Lindberg [29] places his main microphone array (2L-Cube) in the centre of a musical ensemble arranged in a circle to enable the listener to access both sonic and musical details without raising the playback level. Sawaguchi [31] also places the musicians around his middle layer microphone array to provide the listener with a more intimate and engaging experience with the musicians. Such an approach may foster a sensation of social presence, where the listener feels as if the musicians are actually performing in front of him or her.

3.4 Limitations of the Previous Studies and Future Research Required

Previous studies that subjectively compared different 3D microphone arrays, which were reviewed in Sec. 2, are limited in several aspects. Firstly, the number of microphone arrays included in each study was limited to two or three, and in most cases, different microphone models were used for different arrays. The 3D Microphone Array Recording Comparison (3D-MARCo) database [25] provides various types of musical recordings made simultaneously using eight different 3D microphone arrays, six of which were recorded using microphone of an identical brand. Various types of objective measurements of the microphone arrays are provided in [25], but formal subjective evaluations of the recordings are yet to be conducted.

Second, not all studies performed a valid statistical analysis to determine the significance of difference between the arrays tested. Simply comparing the mean averages of the data may lead to a misleading conclusion. Since the result of a subjective listening test can depend not only on the experimental condition, but on the background and experience level of subject group as pointed out in Sec. 2, more careful experimental design for data analysis (e.g., between-subject design as well as within-subject) is required. Furthermore, such factors need to be considered carefully when one interprets the results of 3D audio evaluation from different subject groups. This also means that researchers need to explicitly report the details about their subjects (e.g., training, experience, age, etc.) in their papers.

Third, most of the studies used conventional spatial attributes inherited from the literature on surround sound or/and subjective preference as dependent variables. Kamekawa and Marui [64] elicited 10 attributes from three different microphone arrays, but all of them were related to horizontal spatial impression. Therefore, more complete understanding about salient attributes specific to techniques with different upper microphone layer configurations is required. Several studies reported the existences of vertical attributes such as vertical image spread [12], vertical LEV [87], and engulfment [88]. However, as suggested from [25,64], tonal attributes induced by various types of the upper layer microphone techniques would also be an important aspect to consider in the evaluation of 3D recording techniques.

It is also recognized that microphone techniques for the bottom layer have not yet been formally investigated. Currently, the 9+10+3 and 4+5+1 loudspeaker systems specified in ITU-R BS.2051-2 [5] are the only ones that utilize the bottom layer. However, Grewe et al. [89] experimented with an extension of a 4+5+0 loudspeaker system with three bottom layer loudspeakers, making it a 4+5+3 system. From a listening test that compared the 5+4+0 and other smaller systems against 4+5+3, they found that the 4+5+3 was significantly preferred to the other systems. However, it is not yet clear what kind of low-level perceptual attributes can be enhanced by the addition of negatively elevated loudspeakers and what are the most effective ways to capture sounds for them to reproduce. These topics require systematic investigation.

Finally, most of the studies that compared Ambisonic systems against HVS or HSVC arrays used ITU-R–based, irregular loudspeaker configurations for reproduction. This seems reasonable from the perspective of comparing different microphone systems on a more practical loudspeaker array as an experimental constant. However, as mentioned earlier, a theoretically optimal decoding of an Ambisonic recording require a regular loudspeaker array, while HVS and HSVC arrays are designed for specific irregular loudspeaker arrays. Therefore, it seems difficult to directly compare HVC and HSV/HSVC arrays in a mutually fair way. It is proposed that the subjective qualities of Ambisonic microphone systems need to be evaluated exclusively using a regular loudspeaker array that is ideal for the decoding of their signals.

4 CONCLUSIONS

Existing 3D microphone arrays for music and ambience recording were reviewed in this paper. They are broadly categorised into three groups based on their physical configurations: HVS, HSVC, and HVC. They can be also classified as perceptually motivated and physically motivated arrays. Studies that compared different arrays subjectively or objectively generally suggest that significant differences can be perceived between them in terms of not only spatial attributes but also timbral ones. It also indicates that the perceived qualities of 3D microphone arrays might depend on contextual factors as well as experimental conditions.

The paper also discussed the relative advantages and limitations of different approaches of upper microphone layer configurations, based on psychoacoustic research findings from the literature. The perceptually motivated arrays tend to be split into two groups: those that use upward-facing unidirectional microphones for the upper layer (either vertically coincident or spaced) and those that use vertically spaced omnidirectional microphones. It is considered that the first group would have a more effective vertical channel separation than the second group, leading to a better vertical localization stability and less tonal coloration when recording nonelevated sound sources. This would also allow for a more natural representation of environmental sounds of the recording space (e.g., reflections and reverberation from the ceiling). On the other hand, techniques in the second group might be more useful for creating a vivid sense of height or vertical image spread in reproduction, especially for sound sources that are physically elevated (e.g., pipe organ and choir on risers). They would also be beneficial for achieving a more extended low end in the frequency spectrum. However, care would be required of this approach to avoid an unwanted vertical image shift or a comb-filter effect due to ICXT.

Finally, this paper identified the limitations of the current research in the field of 3D sound recording and topics that need further research. Overall, more rigorous subjective and objective studies are required to define salient perceptual attributes of 3D microphone arrays and their associated objective parameters. This would allow for a more reliable quality evaluation of 3D acoustic recordings. In addition, microphone techniques for the bottom loudspeaker layer should be explored further.

5 ACKNOWLEDGEMENTS

This work was supported by the University of Huddersfield (Grant ref: URF510-01) and Innovate UK (Grant ref: 105175).

6 REFERENCES

[1] Dolby, “Dolby Atmos,” https://www.dolby.com/technologies/dolby-atmos (accessed Aug. 4, 2020).

[2] Auro Technologies, “Auro-3D,” https://www.auro-3d.com (accessed Aug. 4, 2020).

[3] DTS, “Home Theater Sound Gets Real | DTS,” https://dts.com/dtsx (accessed Aug. 27, 2020).

[4] International Telecommunication Union, “Advanced Sound System for Programme Production,” ITU-R Recomm. BS.2051-2 (2018).

[5] International Telecommunication Union, “Multichannel Sound Technology in Home and Broadcasting Applications BS Series,” ITU-R Rep. BS.2159-8 (2019).

[6] J. Herre, H. Johannes, A. Kuntz, and J. Plogsties, “MPEG-H Audio—The New Standard for Universal Spa-

tial/3D Audio Coding," *J. Audio Eng. Soc.*, vol. 62, no. 12, pp. 821–830, (2014 Dec.), doi: 10.17743/jaes.2014.0049.

[7] Dolby, "AC-4 - Dolby Professional," https://professional.dolby.com/technologies/ac-4/ (accessed Aug. 27, 2020).

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

[9] F. Rumsey and W. Lewis, "Effect of Rear Microphone Spacing on Spatial Impression for Omnidirectional Surround Sound Microphone Arrays," presented at the *112th Convention of the Audio Engineering Society* (2002 Apr.), convention paper 5563.

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

[11] H. Lee and C. Gribben, "Effect of Vertical Microphone Layer Spacing for a 3D Microphone Array," *J. Audio Eng. Soc.*, vol. 62, no. 12, pp. 870–884 (2014 Dec.), doi: https://doi.org/10.17743/jaes.2014.004.

[12] C. Gribben and H. Lee, "The Frequency and Loudspeaker-Azimuth Dependencies of Vertical Interchannel Decorrelation on the Vertical Spread of an Auditory Image," *J. Audio Eng. Soc.*, vol. 66, no. 7/8, pp. 537–555 (2018 Jul.), doi: 10.17743/jaes.2018.0040.

[13] M. Gerzon and G. Barton, "Ambisonic Decoders for HDTV," presented at the *92nd Convention of the Audio Engineering Society* (1992 Mar.), convention paper 3345.

[14] F. Zotter and M. Frank, *Ambisonics: A Practical 3D Audio Theory for Recording, Studio Production, Sound Reinforcement, and Virtual Reality*, 1st ed. (Springer, New York, NY, 2019).

[15] M. Gerzon, "General Metatheory of Auditory Localisation," presented at the *92nd Convention of Audio Engineering Society* (1992 Mar.), convention paper 3306.

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

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

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

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

[20] H. Lee, *Effects of Interchannel Crosstalk in Multichannel Microphone Technique*, doctoral thesis, University of Surrey, Guildford, UK (2006).

[21] J. Barbour, "Elevation Perception: Phantom Images in the Vertical Hemi-Sphere," presented at the *AES 24th International Conference on Multichannel Audio* (2003), convention paper 14.

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

[23] R. Wallis and H. Lee, "Vertical Stereophonic Localization in the Presence of Interchannel Crosstalk: The Analysis of Frequency-Dependent Localization Thresholds," *J. Audio Eng. Soc.*, vol. 64, no. 10, pp. 762–770 (2016 Oct.), doi: 10.17743/jaes.2016.0039.

[24] 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.

[25] H. Lee and D. Johnson, "3D Microphone Array Recording Comparison (3D-MARCo): Objective Measurements" (2020), doi: 10.5281/zenodo.4018009.

[26] M. Williams, "Psychoacoustic Testing of the 3-D Multiformat Microphone Array Design and the Basic Isosceles Triangle Structure of the Array and the Loudspeaker Reproduction," presented at the *134th Convention of the Audio Engineering Society* (2013 May), convention paper 9430.

[27] M. Williams, "Microphone Array Design for Localisation With Elevation Cues," presented at the *132nd Convention of the Audio Engineering Society* (2012 Apr.), convention paper 8601.

[28] VDT, "3D Recording with the '2L-cube,'" http://www.2l.no/artikler/2L-VDT.pdf (accessed Aug. 4, 2020).

[29] M. Lindberg, personal communication with the author, 2020.

[30] Merging Technologies, "An Exceptionally Successful 2016 For Unamas And Mick Sawaguchi," https://www.merging.com/news/press-releases/an-exceptionally-successful-2016-for-unamas-and-mick-sawaguchi (accessed Aug. 4, 2020).

[31] M. Sawaguchi, personal communication with the author, 2020.

[32] G. Eskow, "Recording the Orchestra in 9.1," *Mix Magazine*, https://www.mixonline.com/live-sound/recording-orchestra-91-426704 (accessed Aug. 04, 2020).

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

[34] H. Lee, D. Johnson, and M. Mironovs, "An Interactive and Intelligent Tool for Microphone Array Design," presented at the *132nd Convention of the Audio Engineering Society* (2017 Oct.), eBrief 390.

[35] M. Williams and G. Le Du, "Multichannel Microphone Array Design," presented at the *108th Convention of the Audio Engineering Society* (2000 Feb.), convention paper 5157.

[36] H. Wittek, "Image Assistant," https://www.hauptmikrofon.de/stereo-surround/image-assistant (accessed Sep. 15, 2020).

[37] R. Wallis and H. Lee, “The Reduction of Vertical Interchannel Crosstalk: The Analysis of Localization Thresholds for Natural Sound Sources,” *Appl. Sci.*, vol. 7, no. 3, (2017), doi: 10.3390/APP7030278.

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

[39] 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.

[40] 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.

[41] R. K. Cook, R. V. Waterhotjse, R. D. Berendt, S. Edelman, and M. C. Thompson, “Measurement of Correlation Coefficients in Reverberant Sound Fields,” *J. Acoust. Soc. Am.*, vol. 27, no. 6, pp. 1072–1077 (Nov. 1955), doi: 10.1121/1.1908122.

[42] K. Hamasaki, T. Nishiguchi, K. Hiyama, and K. Ono, “Advanced Multichannel Audio Systems With Superior Impression of Presence and Reality,” presented at the *116th Convention of the Audio Engineering Society* (2004 May), convention paper 6053.

[43] 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.

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

[45] H. Lee, “Subjective Evaluations of Perspective Control Microphone Array (PCMA),” presented at the *132nd Convention of the Audio Engineering Society* (2012 Apr.), convention paper 8625.

[46] 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.

[47] M. Williams, “Microphone Arrays for Natural Multiphony,” presented at the *91st Convention of the Audio Engineering Society* (1991 Oct.), convention paper 3157.

[48] 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: 10.17743/jaes.2018.0068.

[49] C. Vaida, “3D Classical Piano Studio Production, Brahms in 13.1 and the au3Dio Microphone Array,” In *Proceedings of the 4th International Conference on Spatial Audio*, pp. 192–196 (2017 Sep.).

[50] E. Sengpiel, “Visualization of All Stereo Microphone Systems with Two Microphones,” http://www.sengpielaudio.com/HejiaE.htm (accessed Sep. 5, 2020).

[51] H. Lee, “The Effects of the Depth and Width of an Ambience Microphone Array on the Perceived Depth, Width and Listener Envelopment,” presented at the *149th Convention of the Audio Engineering Society* (2020 Oct.), convention paper 10434.

[52] A. Avni, J. Ahrens, M. Geier, S. Spors, H. Wierstorf, and B. Rafaely, “Spatial Perception of Sound Fields Recorded by Spherical Microphone Arrays With Varying Spatial Resolution,” *J. Acoust. Soc. Am.*, vol. 133, no. 5, pp. 2711–2721 (2013 May), doi: 10.1121/1.4795780.

[53] E. Kurz, F. Pfahler, and M. Frank, “Comparison of First-Order Ambisonic Microphone Arrays,” presented at the *3rd International Conference on Spatial Audio* (2015 Sep.).

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

[55] M. Frank and F. Zotter, “Exploring the Perceptual Sweet Area in Ambisonics,” presented at the *142nd Convention of the Audio Engineering Society* (2017 May), convention paper 9727.

[56] F. Zotter, H. Pomberger, and M. Noisternig, “Energy-Preserving Ambisonic Decoding,” *Acta Acust. united with Acust.*, vol. 98, no. 1, pp. 37–47 (2012 Jan.), doi: 10.3813/AAA.918490.

[57] F. Zotter and M. Frank, “All-Round Ambisonic Panning and Decoding,” *J. Audio Eng. Soc.*, vol. 60, no. 10, pp. 807–820 (2012 Oct.), doi: 10.1099/ijs.0.63370-0.

[58] P. Geluso, “Capturing Height: The Addition of Z Microphones to Stereo and Surround Microphone Arrays,” presented at the *132nd Convention of the Audio Engineering Society* (2012 Apr.), convention paper 8595.

[59] K. 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.

[60] U. Scuda, H. Stenzel, and H. Lee, “Perception of Elevated Sound Image Recorded with 3D-Audio Microphone Arrays,” presented at the *2nd International Conference on Spatial Audio* (2013 Feb.).

[61] L. Riitano and J. Victoria, “Comparison Between Different Microphone-Arrays for 3D-Audio,” presented at the 1*44th Convention of the Audio Engineering Society* (2018 May), convention paper 9980.

[62] A. Ryaboy, “Exploring 3D: A Subjective Evaluation of Surround Microphone Arrays Catered for Auro-3D Reproduction System,” presented at the *139th Convention of the Audio Engineering Society* (2015 Oct.), convention paper 9431.

[63] W. Howie, D. Martin, D. H. Benson, J. Kelly, and R. King, “Subjective and Objective Evaluation of 9ch Three-Dimensional Acoustic Music Recording Techniques,” in *Proceedings of the 2018 AES International Conference on*

Spatial Reproduction - Aesthetics and Science (2018 Jul.), conference paper P10-1.

[64] 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. Technol.*, vol. 41, no. 1, pp. 260–268 (2020), doi: 10.1250/ast.41.260.

[65] S. Kim, R. King, and T. Kamekawa, "A Cross-Cultural Comparison of Salient Perceptual Characteristics of Height Channels for a Virtual Auditory Environment," *Virtual Real.*, vol. 19, no. 3–4, pp. 149–160 (2015 Aug.), doi: 10.1007/s10055-015-0269-1.

[66] W. Howie, D. Martin, S. Kim, T. Kamekawa, and R. King, "Effect of Audio Production Experience, Musical Training, and Age on Listener Performance in 3D Audio Evaluation," *J. Audio Eng. Soc.*, vol. 67, no. 10, pp. 782–794 (2019 Oct.), doi: 10.17743/jaes.2019.0031.

[67] 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.

[68] F. Rumsey, *Spatial Audio*, 1st ed. (Routledge, Abingdon, UK, 2001).

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

[70] R. King, W. Howie, and J. Kelly, "A Survey of Suggested Techniques for Height Channel Capture in Multi-Channel Recording," presented at the *140th Convention of the Audio Engineering Society* (2016 May), eBrief 266.

[71] S. K. Roffler and R. A. Butler, "Localization of Tonal Stimuli in the Vertical Plane," *J. Acoust. Soc. Am.*, vol. 43, no. 6, pp. 1260–1266 (1968 Jun.), doi: 10.1121/1.1910977.

[72] H. Lee, "Perceptual Band Allocation (PBA) for the Rendering of Vertical Image Spread With a Vertical 2D Loudspeaker Array," *J. Audio Eng. Soc.*, vol. 64, no. 12, pp. 1003–1013 (2016 Dec.), doi: 10.17743/jaes.2016.0052.

[73] R. Wallis and H. Lee, "Localisation of Vertical Auditory Phantom Image With Band-Limited Reductions of Vertical Interchannel Crosstalk," *Appl. Sci.*, vol. 10, no. 4, 1490 (2020), doi: 10.3390/app10041490.

[74] L. Nipkow, "Room Signals – Properties and Influence on Auro 3D Recordings," presented at the *3rd International Conference on Spatial Audio* (2015 Sep.).

[75] W. Howie, R. King, M. Boerum, D. Benson, and A. J. Han, "Listener Preference for Height Channel Microphone Polar Patterns in Three-Dimensional Recording," presented at the *139th Convention of the Audio Engineering Society* (2015 Oct.), convention paper 9419.

[76] M. Kuster, "Spatial Correlation and Coherence in Reverberant Acoustic Fields: Extension to Microphones With Arbitrary First-Order Directivity," *J. Acoust. Soc. Am.*, vol. 123, no. 1, pp. 154–162 (2008), doi: 10.1121/1.2812592.

[77] A. Roginska and P. Geluso, *Immersive Sound: The Art and Science of Binaural and Multi-Channel Audio*, 1st ed. (Routledge, Abingdon, UK, 2017).

[78] L. Ermi and F. Mäyrä, "Fundamental Components of the Gameplay Experience: Analysing Immersion," in *Proceedings of the 2nd Digital Games Research Association International Conference* (2005), pp. 15–27.

[79] E. Brown and P. Cairns, "A Grounded Investigation of Game Immersion," in *CHI '04 Extended Abstracts on Human Factors and Computing Systems*, pp. 1297–1300 (ACM Press, New York, NY, 2004).

[80] H. Lee, "A Conceptual Model of Auditory Immersion in Extended Reality," PsyArXiv (2020), doi: 10.31234/osf.io/sefkh.

[81] F. Biocca, "The Cyborg's Dilemma: Progressive Embodiment in Virtual Environments," *J. Comp. Mediated Commun.* vol. 3, no. 2 (1997), doi: 10.1111/j.1083-6101.1997.tb00070.x.

[82] M. Slater, "Place Illusion and Plausibility Can Lead to Realistic Behaviour in Immersive Virtual Environments," *Philos. Trans. R. Soc. B Biol. Sci.*, vol. 364, no. 1535, pp. 3549–3557 (2009), doi: 10.1098/rstb.2009.0138.

[83] A. Lindau and S. Weinzierl, "Assessing the Plausibility of Virtual Acoustic Environments," *Acta Acust. united with Acust.*, vol. 98, no. 5, pp. 804–810 (2012 Sep.), doi: 10.3813/AAA.918562.

[84] W. Chase, *How Music Really Works!: The Essential Handbook for Songwrighters, Performers and Music Students* (Vancouver, Roedy Black Music, Vancouver, Canada, 2006).

[85] J. Pätynen and T. Lokki, "Concert Halls With Strong and Lateral Sound Increase the Emotional Impact of Orchestra Music," *J. Acoust. Soc. Am.*, vol. 139, no. 3, pp. 1214–1224 (2016), doi: 10.1121/1.4944038.

[86] M. L. Ryan, *Narrative as Virtual Reality: Immersion and Interactivity in Literature and Electronic Media* (The Johns Hopkins University Press, Baltimore, MD, 2003).

[87] H. Furuya, K. Fujimoto, Y. Takeshima, and H. Nakamura, "Effect of Early Reflections From Upside on Auditory Envelopment," *J. Acoust. Soc. Japan*, vol. 2, no. 16, pp. 97–104 (1995), doi: 10.1250/ast.16.97.

[88] R. Sazdov, G. Paine, and K. Stevens, "Perceptual Investigation Into Envelopment, Spatial Clarity, and Engulfment in Reproduced Multi-Channel Audio," in *Proceedings of the 31st International Conference: New Directions in High Resolution Audio* (2007), pp. 1–11.

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

[90] F. Olivieri, N. Peters, and D. Sen, "Scene-Based Audio and Higher Order Ambisonics: A Technology Overview and Application to Next-Generation Audio, VR and 360° Video," https://tech.ebu.ch/publications (2019) (Accessed: Aug. 27, 2020).

THE AUTHOR

Hyunkook Lee

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