OPEN ACCESS Freely available online

Journal of the Audio Engineering Society
Vol. 63, No. 10, October 2015 (© 2015)
DOI: http://dx.doi.org/10.17743/jaes.2015.0075

2D-to-3D Ambience Upmixing Based on Perceptual Band Allocation

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

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

Listening tests were conducted to evaluate the feasibility of a novel 2D-to-3D ambience upmixing technique named "perceptual band allocation" (PBA). Four-channel ambience signals captured in a reverberant concert hall were low-pass and high-pass filtered, which were then routed to lower and upper loudspeaker layers arranged in a 9-channel 3D configuration, respectively. The upmixed stimuli were compared against original 3D recordings made using an 8-channel ambience microphone array in terms of 3D listener envelopment and preference. The results suggest that the perceived quality of the proposed method could be at least comparable to that of an original 3D recording.

0 INTRODUCTION

Three-dimensional multichannel audio systems such as Auro-3D [1], Dolby Atmos [2], and 22.2 [3] employ additional height loudspeakers in order to provide the listener with a three-dimensional (3D) auditory experience. One of the perceptual attributes that could be enhanced by the use of height channels is listener envelopment (LEV). In the context of two-dimensional (2D) surround sound (e.g., 5.1), LEV is widely understood as the subjective impression of being enveloped by reverberant sound [4, 5]. With 3D loudspeaker formats, the added height channels could be used to render the "vertical" spread of reverberant sound image as well as the horizontal one, and ultimately the auditory impression of 3D LEV could be achieved.

One of the key requirements for 3D multichannel audio applications would be a 2D-to-3D upmixing technique that can add a height dimension to 2D content. Therefore, a new method that can render vertical image spread would be necessary. In the context of horizontal stereophony, horizontal image spread can be rendered by means of interchannel decorrelation, and many different decorrelation methods have been proposed over the past years [6–10]. Such methods are based on the principle that as the degree of correlation between stereophonic channel signals decreases, that between ear-input signals (interaural cross-correlation), which has a direct relationship with perceived auditory image spread [4], also decreases. However, vertically reproduced stereophonic signals would have no or little influence on interaural cross-correlation. From a recent study by Gribben and Lee [11] it was found that vertically applied interchannel decorrelation was not as effective as horizontal decorrelation in terms of controlling the spread of image.

The literature generally suggests that vertical localization relies on spectral cues. A number of researchers [12–14] have found that the higher the frequency of a pure tone the higher the perceived image position was regardless of the physical height of the presenting loudspeaker; a phenomenon referred to as the "pitch-height" effect in [15]. In the case of band-pass filtered noise signals, however, this effect was reported to be dependent on the physical height of the loudspeaker that presents the signal. For example, Roffler and Butler [16] found from their experiments using loudspeakers vertically arranged at different heights that the perceived image height of a noise high-pass filtered at 2 kHz was similar to the physical height of the presenting loudspeaker. Conversely, a noise low-passed at 2 kHz was localized around or below the eye level regardless of its presenting loudspeaker height. Similar results were obtained from Cabrera and Tiley's [15] experiment conducted with octave-band noise signals centered at 125 Hz, 500 Hz, 2 kHz, and 8 kHz, using vertically arranged loudspeakers; a higher frequency band was localized at a higher position than a lower frequency band and this difference became larger as the loudspeaker height increased. Cabrera and Tiley [15] and Ferguson and Cabrera [17] confirmed the validity of this phenomenon for low- and high-pass filtered noise stimuli (crossover of 1 kHz) that were simultaneously presented from different loudspeakers at different heights.

The present study aims to explore the feasibility of a new 2D-to-3D upmixing method developed based on the above research findings, which is named "perceptual band allocation" (PBA). The method decomposes the spectrum of the

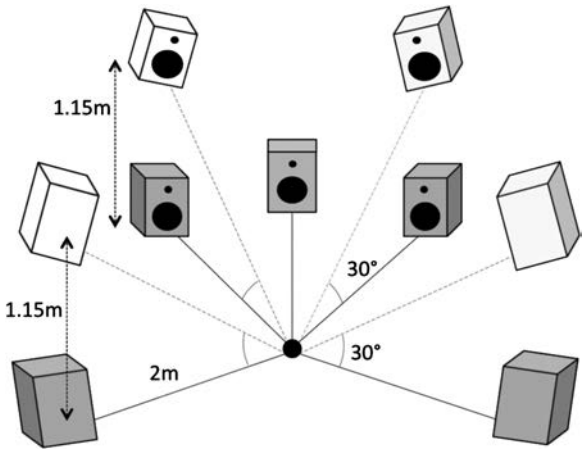

Fig. 1. 9-channel "Auro-3D" reproduction setup used for the listening test.

original signal into sub-frequency bands and maps them to either the lower (main) or upper (height) loudspeaker layer, depending on their unique perceptual positions in the vertical plane, in order to render vertical image spread. In the current experiment, simple 2-band PBA scenarios have been tested in the context of the 2D-to-3D upmixing of ambience. The low-pass and high-pass filtered frequency contents of original ambience signals recorded in a concert hall were allocated to main and height channels in a 9-channel Auro-3D loudspeaker format, with three different crossover frequencies. Listening tests were conducted to examine the effectiveness of this method for the perception of 3D LEV and subjective quality preference.

The rest of this paper consists of the following. Sec. 1 describes the experimental method. Sec. 2 presents the statistically analyzed results of the listening tests conducted. Sec. 3 discusses the results and describes further works. Finally, Sec. 4 summarizes and concludes the work.

1 EXPERIMENTAL METHOD

1.1 Reproduction Format

The loudspeaker configuration used for the current experiment was based on the Auro-3D 9-channel format [1], as shown in Fig. 1. A total of nine loudspeakers were set up in a dry listening room (8.3m (W) x 5.4m (L) x 3.4m (H); RT = 0.2s) at the University of Huddersfield. Five Genelec 8040A loudspeakers were situated in a conventional 5-channel arrangement [18], with the azimuths of the front left and right loudspeakers from the center loudspeaker being $\pm 30^\circ$ and those of the rear left and right being $\pm 120^\circ$. An upper layer of four height-channel loudspeakers of the same type was placed directly above the main layer, at a vertical angle of 30° from the listening position. ITU-R BS. 2051 [19] recommends that the height channel elevation angle should be between 30° and 45°. The distance from each loudspeaker to the listening position was 2 m. The loudspeakers were aligned in time delay and sound pressure level (SPL) at the listening position.

1.2 Recording

1.2.1 Physical Setup

In order to create test stimuli, multichannel room impulse responses (MRIRs) and a virtual string quartet were recorded in a reverberant concert hall called St. Paul's in Huddersfield, UK (V = approx. 5700m^3; RT = avg. 2.1s). The recording setup is shown in Fig. 2. Five Genelec 8040A loudspeakers were placed on the stage, at a 1 m height from the stage floor. The MRIRs were acquired with the center loudspeaker using the exponential sine sweep method described in [20].

1.2.2 Microphone Techniques

For source imaging, an "ICA-3" 3-channel frontal microphone array [21] was placed 2.5 m away from the center loudspeaker and raised 2.2 m high from the stage floor. This array has been reported to provide sufficient frontal spatial impression [22] as well as a continuous image localization across left, center, and right loudspeakers [21]. The signals captured by the microphones—L, C, and R—were to be routed to the front three loudspeakers. The microphone array used for capturing diffused ambience was a "Hamasaki-Square" (HS) [23], which employs four side-facing figure-of-eight microphones arranged in a 2 m $\times$ 2 m square. Two-m was recommended as an optimal microphone spacing to produce LEV based on a theoretical calculation showing that this spacing provides sufficient interchannel decorrelation above 100 Hz for two omni-directional microphones in a diffuse field. The front two microphones (FL and FR) were to feed the front left and right loudspeakers, while the rear two (RL and RR) the rear left and right loudspeakers. This technique, blending the ambience from the front with that from the rear, has been reported to produce a greater LEV than 2-channel rear microphone techniques [22]. The front microphone pair of the HS was placed 10 m away from the center loudspeaker, which is beyond the critical distance of the venue; this was to ensure negative values of direct to reverberant (D/R) energy ratio of the recorded signals.

In order to create reference 3D ambience signals, which were to be compared against PBA-upmixed signals in the listening tests, a new microphone technique was devised; the original HS was augmented with four upward-facing cardioid microphones: FLh, RFh, RLh, and RRh. The height microphones were placed 1 m directly above the HS, and angled vertically so that they could capture ambient sounds mainly from the ceiling. At the point of the current experiment in time, there was no microphone array configuration specifically developed for capturing 3D ambience. However, the microphone configuration devised for the current experiment was found to be similar to that recently proposed by Hamasaki and Baelen [24] in that both use ceiling-facing cardioid microphones placed directly above a base HS layer. The polar pattern and positioning aspects of height ambience microphone will be further discussed in Sec. 3.3.

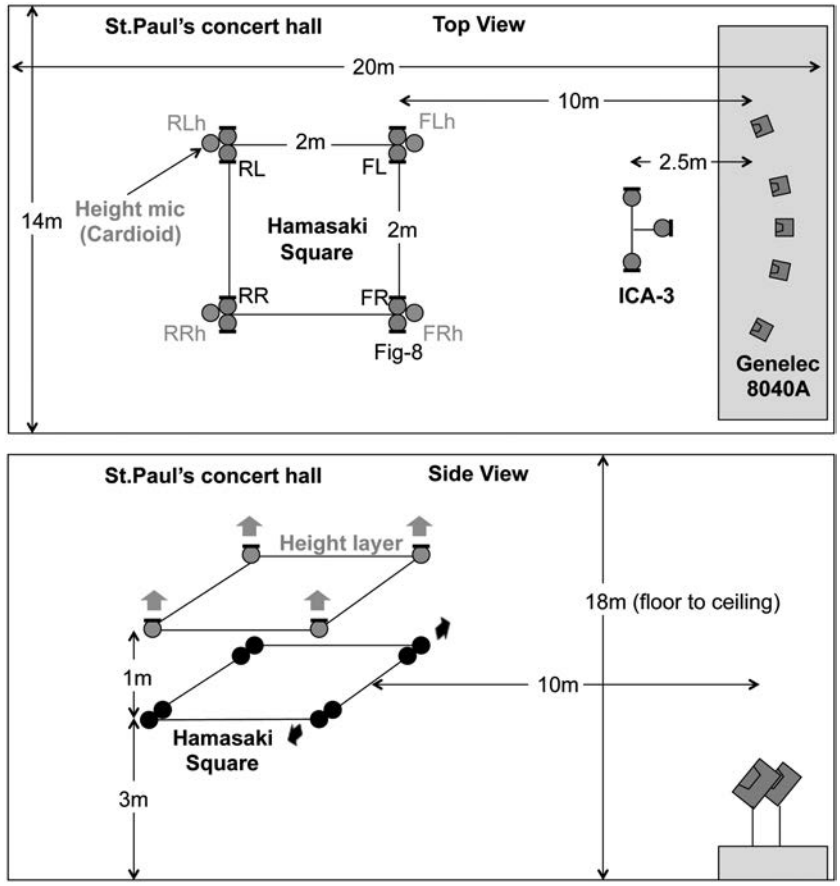

Fig. 2. Recording setup.

Table 1. Direct to reverberant (D/R) energy ratio for the room impulse responses captured by the ambience microphones.

Channel	FL	FR	RL	RR	FLh	FRh	RLh	RRh
D/R ratio (dB)	–8.5	–8.0	–10.0	–9.5	–4.5	–5.3	–5.8	–5.7

Table 2. Interchannel cross-correlation coefficients (ICCCs) for the main-height microphone pairs of FR-FRh and RR-RRh; calculated for octave-bands and averaged for low, mid, and high bands.

	Time segment	Low band average (63, 125, 250 Hz)	Mid band average (500, 1 k, 2 kHz)	High band average (4 k, 8 k, 16 kHz)
FR-FRh	Early (0. . .80ms)	0.60	0.32	0.30
	Late (80. . .750ms)	0.33	0.13	0.10
RR-RRh	Early (0. . .80ms)	0.69	0.35	0.24
	Late (80. . .750ms)	0.29	0.14	0.10

1.2.3 Signal Relationship

The MRIRs were used to analyze the direct to reverberant (D/R) energy ratio of each channel signal (Table 1) as well as interchannel cross-correlation coefficients (ICCC) (Table 2). The D/R energy ratio, defined as Eq. (1) in the Appendix, provides an indication as to how much suppression of direct sound was achieved in the signals. The ICCC, which is defined as Eq. (2) in the Appendix, is used here for measuring the degree of channel separation.

1.3 Stimuli Creation

The MRIRs were convolved with anechoic trumpet and conga recordings to create single source stimuli. These sources were chosen to examine the influences of different temporal and spectral characteristics. The other four loudspeakers were arranged in a string quartet formation with 4 m width to reproduce a 25-second excerpt from Vivaldi's "Four Seasons: Summer," which was programmed by a professional orchestral music arranger using the Westwood virtual orchestra software. This piece was chosen for its complex temporal characteristics as well as a broad frequency spectrum. The waveforms and spectra of the sound sources used are shown in Fig. 3.

For the 2-band PBA upmixing, each of the ambience signals captured by the main HS array was first split into low and high bands at three different crossover frequencies of 0.5 kHz, 1 kHz, and 4 kHz (rolled off at 48 dB/octave). The low and high bands were then fed into the main and height loudspeakers, respectively (Fig. 4). The variation of crossover frequency was to examine the effect of allocating different amounts of high and low frequencies to the main and height speakers.

From the above process, a total of seven stimuli were created for each of the trumpet, conga, and string quartet sources. The routing scheme for each stimulus type

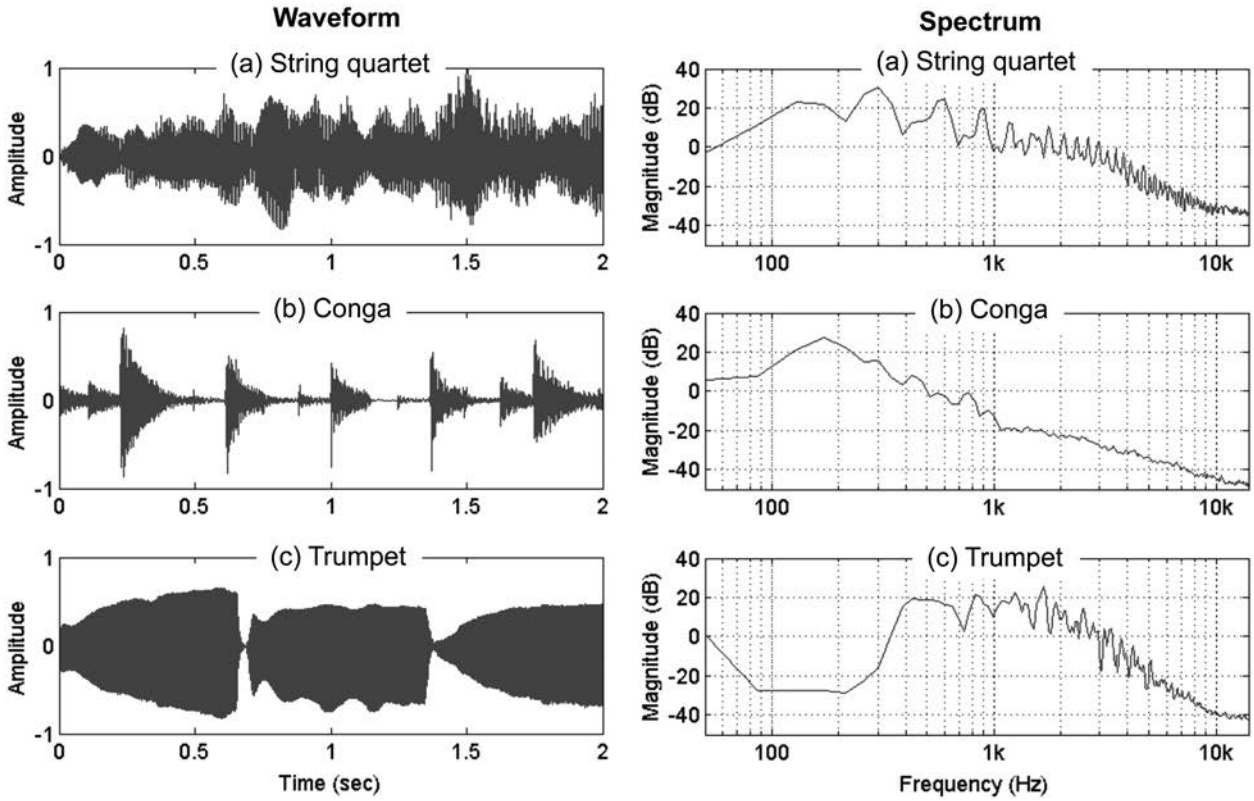

Fig. 3. Waveforms and spectra of stimuli created: signals captured by the front center channel of the frontal microphone array used.

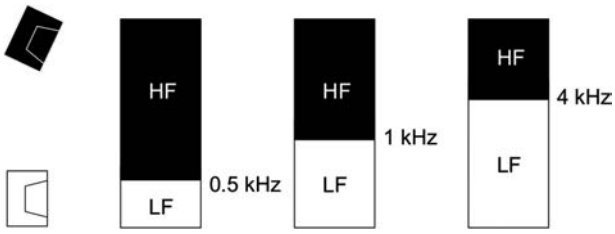

Fig. 4. 2-band decomposition of original ambient signals at three different division frequencies; black and white areas represent frequency components allocated to the height and main channels, respectively.

Table 3. Descriptions of routing scheme used for the listening test stimuli

Stimuli	Descriptions of routing scheme
3-0-0	Only ICA-3 reproduced from the front three loudspeakers
3-2-0	3-0-0 with the 4-channel HS ambience signals reproduced from the left and right main layer loudspeakers in the front and rear
3-0-4	3-0-0 with the HS signals reproduced from the left and right height layer loudspeakers in the front and rear
3-2-4	3-2-0 with the additional four height microphone signals reproduced from the height layer loudspeakers
PBA-0.5k	PBA-upmixed version of 3-2-0 with 0.5 kHz Fc
PBA-1k	PBA-upmixed version of 3-2-0 with 1 kHz Fc
PBA-4k	PBA-upmixed version of 3-2-0 with 4 kHz Fc

is named and described in Table 3. The conventional 5-channel surround format is often referred to as "3-2" format as the format uses three front and two rear loudspeakers. Similarly, the general naming fashion used in this paper follows the number of loudspeakers used for the front main, rear main, and all of the height channels, e.g., 3-2-4 means four height loudspeakers are utilized together with the conventional 5-channel loudspeakers, thus nine channels in total.

In the evaluation of 2- to 5-channel ambience upmixing, upmixed versions are often compared only against one reference, which is their 2-channel original recording, on a multiple stimulus bipolar scale [25, 26]. While this method is valid for ranking different upmixed versions against the original recording, it is not possible to examine how optimal the quality of each version is due to the absence of a high-quality 5-channel anchor in the stimuli set. Arguably, a high anchor for upmixing evaluation would be an original 5-channel recording with rear channel ambience signals that are sufficiently decorrelated and well-balanced with source signals, e.g., multichannel recording made in a concert hall using a separate ambience microphone array, or pop music mix with artificial reverberation fed into rear channels. Similarly, a suitable high anchor for 2D-to-3D ambience upmixing would be a high quality original 3D ambience recording, which could also be used as a reference for aligning the levels of upmixed ambience signals. In the current 3D LEV experiment, therefore, the 3-2-4 stimulus, which is the original 3D recording made using the 8-channel ambience microphone array described above, served as a

high hidden anchor. The 5- and 3-channel stimuli 3-2-0 and 3-0-0 were included as middle and low hidden anchors, respectively. The 3-0-4 stimulus was to test a condition where the original full-band ambience signals captured by the main layer of the ambience array were mapped to the height loudspeakers instead of the main ones. This condition could also be regarded as a PBA condition with the crossover frequency of 0 Hz.

1.4 Playback Level Alignment

The average SPL (*Leq*) of each upmixed stimulus involving height channels was aligned as reference to that of the 3-2-4, which was 78 dB(A), at the listening position. For this, only the levels of ambience signals were changed, while those of the front ICA-3 signals were kept constant. The original level balance between the low and high band ambience signals for PBA was also maintained during the level calibration. The *Leq*s for the 3-0-0 and 3-2-0 were 72.3 dB(A) and 74.9 dB(A), respectively. This shows that the addition of the extra four height signals to the original HS caused a 3.1 dB increase in overall SPL at the listening position. This also means that the levels of all the PBA upmixed stimuli were 3.1 dB higher than the 3-2-0. This kind of increase in level is a typical result in upmixing due to the use of extra channels, as can be observed also in [27].

1.5 Test Method

Fourteen subjects took part in both the 3D LEV and preference listening tests. They were selected from the staff members, research students, and final year music technology students at the University of Huddersfield. All of them had previous experience in multichannel spatial quality evaluation. The test order was randomized for each subject.

In the 3D LEV tests, the subjects were asked to compare and grade all of the seven stimuli listed in Table 3, for each sound source in each trial, in terms of the perceived magnitude of 3D LEV. They were provided with a graphical user interface (GUI), which was developed by the author using the Max-MSP software. The stimuli were synchronized and played back in loop for simultaneous comparison, and the subjects were allowed to listen to them as many times as they wanted. A continuous grading scale ranging from 0 to 100 was used. The directions of grading were indicated as "greater" towards 100 and "lesser" towards 0. The subjects were instructed to interpret 3D LEV as a global attribute that describes the auditory sensation of being surrounded by reverberant sound vertically as well as horizontally. Prior to the main tests, they were given a familiarization trial with broadband male speech stimuli for all conditions under testing. The presentation orders of the stimuli and trials were randomized for each subject in order to avoid potential psychological biases.

The preference tests were conducted using the same GUI and grading scale. However, the 3-2-0 and 3-0-0 stimuli were excluded in these tests so that the subjects could focus solely on perceptual differences among the 3D stimuli only. Furthermore, it was considered unsuitable to compare among 1D, 2D, and 3D sounds in terms of preference since differences in perceived loudness might bias the subjects' preference judgments. In addition to the preference grading, each subject was asked to freely describe attributes that most predominantly affected his or her preference judgments for each sound source.

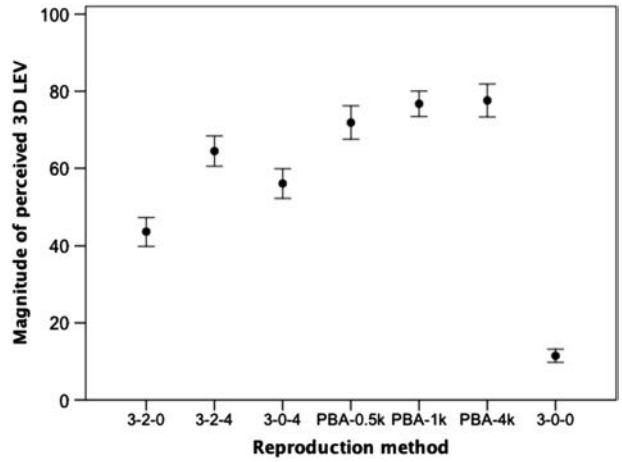

Fig. 5. Mean values and the associated 95% confidence intervals of the 3D LEV data for all sources.

2 RESULTS

Data collected from the listening tests were first normalized according to the recommendation of ITU-R BS.1116-2 [28], and statistically analyzed using the SPSS software. Shapiro-Wilk and Levene's tests confirmed that the data for each reproduction method met the assumption of normal distribution and equal variance for parametric testing. Repeated Measure (RM) ANOVA tests were performed to examine the main effects of reproduction method and sound source type, and post-hoc multiple comparison tests were carried out to compare differences among all. For the multiple comparison results, the Bonferroni correction was applied to the p values in order to avoid potential type-I errors.

2.1 3D Listener Envelopment

The RM ANOVA indicated that the main effect of reproduction method for 3D LEV was significant ($p < 0.01$, $F = 229.882$), although that of source was not ($p > 0.05$). Fig. 5 plots the mean values and the associated 95% confidence intervals for all sources. It can be first observed that all 3D reproduction methods were graded higher than the 2D ones, which was an expected result. More interestingly, all the PBAs appear to be higher than 3-2-4 as well as 3-0-4. The Bonferroni-corrected multiple comparison tests revealed a significant difference between 3-2-4 and PBA-1k or PBA-4k. However, there was no significant difference found between any of the PBA stimuli.

Although the source effect for 3D LEV was not significant overall, the RM ANOVA showed that the interaction between source and reproduction method was significant for 3D LEV ($p < 0.01$, $F = 3.694$). This can be observed

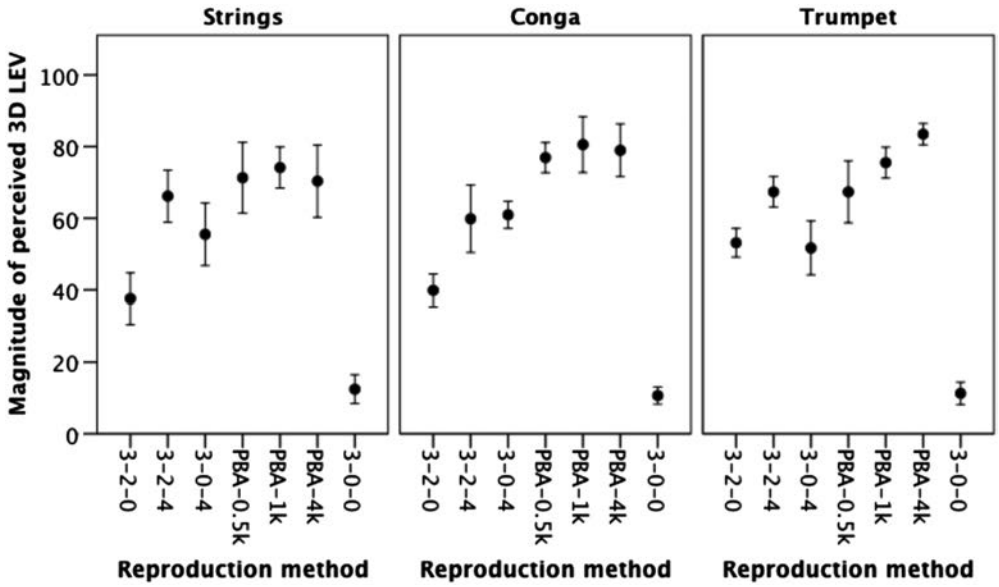

Fig. 6. Mean values and the associated 95% confidence intervals of the 3D LEV data for each source.

visually in Fig. 6, which plots the results for each source separately. There are three main interaction aspects to be considered. First, the trumpet results show a linear increasing pattern for the PBAs; PBA-4k produced the largest 3D LEV, followed by PBA-1k and PBA-0.5k in order. The multiple comparison tests suggest that the difference between PBA-0.5k and PBA-4k and that between PBA-1k and PBA-4k were significant at the 5% level. The other two sources did not show any significant differences among the PBA results. Second, the difference between 3-2-4 and the PBAs of different crossover frequencies depended on sound source. For the conga, the PBA produced a significantly greater 3D LEV than 9ch original ($p < 0.01$), regardless of its crossover frequency. PBA-1k and PBA-4k for the trumpet were also significantly greater than 3-2-4 ($p < 0.01$), whereas PBA-0.5k was not ($p > 0.05$). For the strings, however, there was no significant difference observed between 3-2-4 and any of the PBAs ($p > 0.05$). Finally, 3-2-0 was graded similarly to 3-0-4 for the trumpet, although the former had a lower level of ambience than the latter. For the other two sources, on the other hand, 3-0-4 produced a significantly greater 3D LEV than 3-2-0.

2.2 Preference

It was found from the RM ANOVA that the main effect of reproduction method was significant for subjective preference ($p < 0.01$, $F = 9.690$), whereas neither the main effect of source nor the interaction between method and source was significant ($p > 0.05$). Therefore, the data for different methods combined for all sources are plotted in Fig. 7. As can be seen, all the PBA stimuli were preferred to both 3-2-4 and 3-0-4. The Bonferroni-corrected multiple comparison tests confirmed that this difference was statistically significant at the 5% level, and also that the different crossover frequencies for PBA did not cause any significant difference in subjective preference.

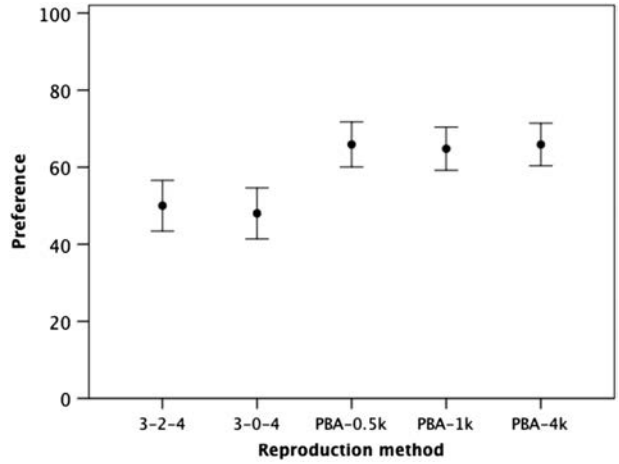

Fig. 7. Mean values and the associated 95% confidence intervals of the preference data for all sources.

Although overall source-method interaction was found to be insignificant, the dependency of the difference between the PBA and 3-2-4 stimuli on the source type is revealed from analyzing the results for each source separately. From the multiple comparison tests for individual sources, it was found that the 3-2-4 and any of the PBAs did not have a significant difference for the strings and trumpet, whereas the 3-2-4 was graded significantly lower than all PBAs for the conga. This can also be observed visually in Fig. 8.

Table 4 shows the list of preference attributes collected from the subjects. The numbers in the brackets indicate the number of occurrences. Through informal discussions with test subjects, synonyms were grouped by the author into common terms. Overall, the table shows that the weighting between spatial and timbral attributes tends to vary depending on the sound source. For example, for the conga, 8 and 11 responses were given to the spatial and timbral aspects, respectively, whereas for the strings and trumpet the

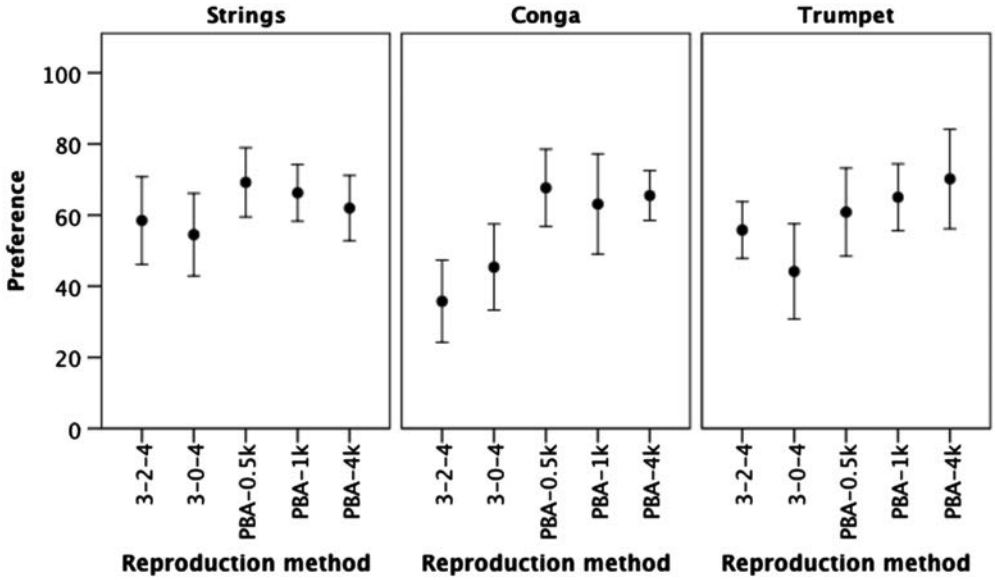

Fig. 8. Mean values and the associated 95% confidence intervals of the preference data for each source.

Table 4. Preference attributes elicited from the subjects

Source	Spatial attributes	Tonal attributes
Strings	Envelopment [7]	Tonal balance [5]
	Spatial naturalness [1]	Clarity [2]
		Fullness [1]
Conga	Envelopment [5]	Boominess [6]
	Localizability [1]	Clarity [3]
	Depth [1]	Tonal balance [1]
	Spatial naturalness [1]	Tonal naturalness [1]
Trumpet	Envelopment [7]	Clarity [7]
	Depth [1]	Tonal balance [1]
	Spatial naturalness [1]	Boominess [1]

weighting was the same. Furthermore, the type of predominant tonal attribute also differs for different sources, although "envelopment" was the most frequently mentioned spatial attributes for all sources. For instance, the main preference attribute for the conga was the strong resonance that occurred at low-mid frequencies, which was a negative factor. Conversely, those for the strings and trumpet were "tonal balance" and "clarity," respectively.

3 DISCUSSION

3.1 3D Listener Envelopment

3.1.1 Original 3D vs. PBA-Upmixed 3D in Vertical LEV

The PBA-upmixed versions of the original 2D recordings were found to be similar to or greater than the 3-2-4 stimuli (original 3D recordings) in the perceived magnitude of 3D LEV, depending on the crossover frequency. Possible explanations for this result are given as follows. For the 3-2-4, both main and height loudspeaker layers presented ambience signals with similar spectral contents, thus producing a vertical phantom image. In the case of horizontal stereophony, the perceived image spread of a phantom image can be increased effectively by de-correlating the loudspeaker signals. However, the results of [11, 29] suggest that vertical interchannel decorrelation is not effective for increasing vertical image spread. Therefore, although the main and height channel signals of the 3-2-4 had sufficiently low ICCCs at mid and high frequencies as shown in Table 2, vertical image perceived for the 3-2-4 condition might have not been fully spread across the positions of the main and height loudspeakers.

On the other hand, the PBA stimuli had no overlapping of frequency contents between the main and height channels, and their perceived vertical image spreads are purely based on the pitch height effect. Frequency components presented from a loudspeaker inherently have a vertical displacement of their perceived images, and the range of this displacement becomes larger as the physical height of the loudspeaker increases [15, 17]. Research also suggests that frequencies above around 6 kHz are localized accurately at the physical height of the loudspeaker, whereas the low frequencies below 1 kHz are localized around the listener's ear height regardless of the loudspeaker's physical height [16]. From this, it is considered that when presenting low and high frequencies independently from the main and height loudspeaker layers, respectively, it is possible to achieve a full vertical spread of the entire image.

3.1.2 PBA Crossover Frequency and Horizontal LEV

The result showing that the 3-0-4 stimuli produced a smaller 3D LEV than the PBA stimuli seems to suggest the importance of the presence of main layer signals for the perception of 3D LEV in multichannel reproduction. That is, even if broadband ambience signals presented by the height channels could contribute to the perception of vertical LEV owing to the pitch-height effect, horizontal LEV produced directly by the main channels would still play an important role on the perception of overall LEV.

In relation to the above point, it is worth noting that the effect of the crossover frequency on the 3D LEV results depended on the type of sound source. For the trumpet, the magnitude of perceived 3D LEV increased almost linearly with a statistical significance as the crossover frequency increased, whereas the results for the strings and conga sources were hardly affected by the crossover frequency. This seems to be associated with the bandwidth of source signal and its influence on horizontal LEV produced by the main loudspeaker layer. As shown in Fig. 3, the lowest frequency of the trumpet signal was around 400 Hz. This means that the PBA-0.5k stimulus had little signal presented from the main loudspeaker layer, while almost all frequencies were presented from the height layer. On the other hand, the main layer signals of the PBA-4k contained frequencies between around 400 Hz and 5.7 kHz, while there was little energy from the height layer due to the high frequency roll-off characteristics of the original source spectrum. Hidaka et al. [4] suggest that the effect of change in interaural cross-correlation for the perceptions of horizontal spatial impression is greatest for the octave-bands centered at 500 Hz, 1 kHz, and 2 kHz, which the PBA-0.5k for the trumpet lacked in the main channels (i.e., $IACC_{E3}$ for apparent source width (ASW) and $IACC_{L3}$ for LEV). From this it is considered that the result for the trumpet source was mainly influenced by the perceived magnitude of horizontal LEV rather than that of vertical LEV. This further suggests that the crossover frequency for the PBA should be determined adaptively based on the spectral characteristics of sound source.

3.2 Preference

All of the PBA-upmixed stimuli were generally found to be preferred to the original 3D stimuli, and this was most obvious for the conga source. As presented in Table 4, "boominess" in the original 3D recording was the most frequently answered reason for the subjects' preference judgments. This seems to be related to the temporal and spectral characteristics of the sound source. First, since the conga is a transient source, early reflections and reverberation produced by the source would be more clearly audible than those produced by a more continuous source. Second, as shown in Fig. 3, the frequency spectrum of the conga has a dominance at around 180 Hz. Furthermore, the ICCCs between the main and height channel signals for the low frequency components were higher than those for the mid and high components (Table 2). Therefore, when the ambient signals reproduced from four pairs of vertically arranged loudspeakers were summed at the listener's ears, the low frequency components of the signals would have been perceptually emphasised more than the mid and high components, thus causing the "boominess." On the other hand, the PBA stimuli have no overlapping of frequencies at the ear since each frequency band is allocated to either the main or height loudspeaker layer independently, and therefore no particular frequencies are boosted in level.

Additionally, the current preference results seem to suggest the importance of timbral quality as well as that of spatial quality in 3D sound recording. This is supported by the fact that the subjects' main reasons for preferences were almost equally divided into timbral and spatial attributes for all sources. Further research is required on this topic.

3.3 Limitations and Future Works

It should be noted that the present result might have a dependency on the microphone technique used and the acoustic condition of the recording venue used. Therefore, it is only tentatively concluded here that the PBA method could produce a sound quality that is at least comparable to that of an original 3D recording. In fact, the optimal way of capturing height information for 3D sound is an area that needs further psychoacoustic experiments. In the current experiment, four upward-facing cardioids were employed for height microphones in order to mainly capture ambience from the ceiling, which is similar to an approach proposed in [24]. However, due to the polar pattern and the direction of the microphones, the D/R ratios of the height signals were about 3–4 dB higher than those of the sideward-facing figure-of-eight microphone signals of the main layer as presented in Table 1. In order to maximally lower D/R ratios of height channel signals, the cardioid microphones could be positioned so that their null-points face towards sound source. Alternatively, side-facing or up-down-facing figure-of-eight could be employed in order to reduce direct sound components in the captured signals. These ideas will be tested in various acoustic environments in a future study.

For a more conclusive evaluation of the PBA upmixing, a wide variety of sound sources including a large scale orchestra as well as single sources are to be tested. The scope of the current work was limited to the 5-channel (2D) to 9-channel (3D) upmixing of ambience signals captured in a real acoustic space using a microphone array. Therefore, source and ambience were treated separately, and this is most relevant for upmixing at a content creation stage. In the case of a "blind" 3D upmixing of 2D content using a commercial AV receiver, on the other hand, the PBA could be applied to ambient components extracted from the original content using such a method as Principal Component Analysis (PCA). In addition, although the PBA is considered unsuitable to be applied for source images that need accurate localization in the vertical plane, it could still be useful for applications where the main requirement is creative rendering of a vertically diffused source image rather than accurate localization, e.g., spatialization in electroacoustic composition.

The present study only considered simple 2-band PBA scenarios in order to examine the feasibility of the method for 2D-to-3D upmixing. Although the 2-band method described in this paper could be applied for practical applications directly, the original aim of the PBA method is to render different degrees of perceived vertical image spread by exploiting the "pitch-height" effect. For this, an original signal is decomposed into multiple frequency bands, which are then allocated to either main or height channels independently depending on their unique perceptual positions in the vertical plane. This work will be conducted for

various loudspeaker azimuth angles, in both real and phantom image conditions, and results will be applied to the practical PBA upmixing evaluations mentioned above.

Last, the present study tested a global attribute of 3D LEV. A future study will investigate the relative perceptual weighting between horizontal LEV and vertical LEV on overall 3D LEV and its dependency on the characteristics of sound source. Further to this, more detailed perceptual attributes of 3D sound recording will be elicited and dimensionalized through formal experiments.

4 SUMMARY AND CONCLUSIONS

This paper introduced a new 2D-to-3D upmixing method named "perceptual band allocation" (PBA), which is based on a psychoacoustic principle of vertical sound localization, the so called "pitch height" effect. The practical feasibility of the method was investigated using 4-channel ambience signals recorded in a reverberant concert hall using the Hamasaki-Square (HS) microphone technique. The original ambience signals were split into lower and upper frequency bands at three crossover frequencies of 0.5 kHz, 1 kHz, and 4 kHz. The loudspeaker setup used was based on Auro-3D 9-channel configuration (four height loudspeakers placed directly above the front left, front right, rear left, and rear right main loudspeakers and elevated at 30° from the listener's ear level). The high-passed signals were fed to the height loudspeakers, while the low-passed ones were routed to main loudspeakers directly below the height loudspeakers. The front left, center, and right main loudspeakers were fed by signals recorded with an ICA-3 microphone array. Three sound sources comprising the recordings of string quartet, conga, and trumpet were used. Multiple comparison tests were conducted on a continuous scale to grade the PBA-upmixed stimuli against 9-channel 3D recordings made using the ICA-3 and the HS augmented with four extra height microphones, in terms of two global attributes: 3D listener envelopment (LEV) and preference. Other stimuli included in the tests were the recordings of the 3-channel ICA-3, 5-channel 2D mix of ICA-3 and HS, and 7-channel 3D mix of ICA-3 and HS.

Data from the tests were statistically analyzed. Results showed that the PBA-upmixed 3D stimuli were significantly greater than or similar to the 9-channel 3D stimuli in 3D LEV, depending on the sound source and the crossover frequency of PBA. They also significantly produced greater 3D LEV than the 7-channel 3D stimuli. Although the result might have been dependent on the microphone techniques used and the acoustics of the recording venue, this result seems to suggest that the spatial quality of PBA-upmixing is at least comparable to that of a real 3D recording. It was also found that there was a significant interaction between source and PBA crossover frequency. For the trumpet, which lacked energies below 400 Hz, the magnitude of perceived 3D LEV increased almost linearly as the crossover frequency increased. The above results suggest the dependency of PBA on the spectral characteristics of sound source and the importance of sound radiated from horizontal plane loudspeakers for the perception of 3D LEV. For the preference tests, the PBA stimuli were significantly preferred to the original 9-channel stimuli overall. For the conga source in particular, the most salient attribute for the preference judgment was "boominess" perceived with the original 3D recording rather than any spatial attributes. This implies that a negative tonal coloration could occur when multiple signals of similar frequencies are combined at the ears from the main and height loudspeakers, depending on the spectral and temporal characteristics of sound source and the recording techniques used. The PBA method, on the other hand, could be beneficial in that the spectrum of original signal can be reconstructed at the ear without any comb-filtering.

5 ACKNOWLEDGMENT

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC), UK, Grant Ref. EP/L019906/1. The author thanks the music technology students and staff at the University of Huddersfield who participated in the listening tests. He is also grateful to the editor and anonymous reviewers of this paper for their insightful and constructive comments.

6 REFERENCES

[1] W. V. Baelen and B. Claypool, Auro 3D whitepaper, http://www.auro-3d.com/wp-content/uploads/2011/05/Whitepaper-Auro3D-Barco.pdf.

[2] Dolby, "Dolby Atmos Next Generation Audio for Cinema," http://www.dolby.com/us/en/technologies/dolby-atmos/dolby-atmos-next-generation-audio-for-cinema-white-paper.pdf

[3] ITU-R, "Report ITU-R BS.2159-4 Multichannel Sound Technology in Home and Broadcasting Applications," *International Telecommunications Union* (2012).

[4] T. Hidaka, L. Beranek, and T. Okano "Interaural Cross-Correlation Lateral Fraction, and Low- and High-Frequency Sound Levels as Measures of Acoustical Quality in Concert Halls," *J. Acoust. Soc. Am.*, vol. 98, pp. 988–1007 (1995), http://dx.doi.org/10.1121/1.414451

[5] J. Bradley and G. Soulodre, "Objective Measures of Listener Envelopment," *J. Acoust. Soc. Am.*, vol. 98, pp. 2590–2597 (1995), http://dx.doi.org/10.1121/1.413225

[6] H. Lauridsen, "Experiments Concerning Different Kinds of Room Acoustics Recording," *Ingenioren*, pp. 47 (1954).

[7] G. S. Kendall. "The Decorrelation of Audio Signals and Its Impact on Spatial Imagery," *Computer Music J.*, vol. 19, pp. 71–87 (1995). http://dx.doi.org/10.2307/3680992

[8] G. Potard and I. Burnett "Decorrelation Techniques for the Rendering of Apparent Sound Source Width in 3D Audio Displays," in *Proc. DAFx-04*, pp. 280–284 (2004).

[9] F. Zotter, and M. Frank, "Efficient Phantom Source Widening," *Arch. Acoust.*, vol. 38, pp. 27–37 (2013), http://dx.doi.org/10.2478/aoa-2013-0004

[10] T. Pihlajamäki, O. Santala and V. Pulkki, "Synthesis of Spatially Extended Virtual Source with Time-Frequency Decomposition of Mono Signals,"

J. Audio Eng. Soc., vol. 62, pp. 467–484 (2014 Jul./Aug.), http://dx.doi.org/10.17743/jaes.2014.0031

[11] C. Gribben and H. Lee, "The Perceptual Effects of Horizontal and Vertical Interchannel Decorrelation Using the Lauridsen Decorrelator," presented at the *136th Convention of the Audio Engineering Society* (2014 Apr.), convention paper 9027.

[12] C. C. Pratt, "The Spatial Character of High and Low Tones," *J. Exp. Psychol.*, vol. 13, pp. 278–285 (1930).

[13] O. C. Trimble, "Localization of Sound in the Anterior–Posterior and Vertical Dimensions of 'Auditory' Space," *Brit. J. Psychol.*, vol. 24, pp. 320–334 (1934).

[14] S. K. Roffler and R. A. Butler, "Localization of Tonal Stimuli in the Vertical Plane," *J. Acoust. Soc. Am.*, vol. 43, pp. 1260–1266 (1968), http://dx.doi.org/10.1121/1.1910977

[15] D. Cabrera and S. Tilley, "Vertical Localization and Image Size Effects in Loudspeaker Reproduction," presented at the *AES 24th Int. Conf.: Multichannel Audio: The New Reality* (2003 Jun.), conference paper 46.

[16] S. K. Roffler and R. A. Butler, "Factors that Influence the Localization of Sound in the Vertical Plane," *J. Acoust. Soc. Am.*, vol. 43, pp. 1255–1259 (1968), http://dx.doi.org/10.1121/1.1910976

[17] S. Ferguson and D. Cabrera, "Vertical Localization of Sound from Multiway Loudspeakers," *J. Audio Eng. Soc.*, vol. 53, pp. 163–173 (2005 Mar.).

[18] ITU-R, "Recommendations ITU-R BS.775-3 Multichannel Stereophonic Sound System with and without Accompanying Picture," *International Telecommunications Union* (2012).

[19] ITU-R, "Recommendations ITU-R BS. 2501-0 Advanced Sound System for Programme Production," *International Telecommunications Union* (2014).

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

[21] U. Herrmann and V. Henkels, "Main Microphone Techniques for the 3/2-Stereo- Standard," in *Proc. 20th Tonmeistertagung* (1998).

[22] R. Kassier, H-K. 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), convention paper 6429.

[23] K. Hamasaki, "Multichannel Recording Techniques for Reproducing Adequate Spatial Impression," presented at the *AES 24th Int. Conference: Multichannel Audio, The New Reality* (2003 Jun.), conference paper 27.

[24] K. Hamasaki, and W. V. 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.

[25] T. Sporer, A. Walther, J. Liebetrau, S. Bube, C. Fabris, T. Hohberger, and A. Köhler, "Perceptual Evaluation of Algorithms for Blind Up-Mix," presented at the *121st Convention of the Audio Engineering Society* (2006 Oct.), convention paper 6915.

[26] H. Purnhagen, A. Ehret, J. Rödén and A. Gröschel, "A Novel Approach to Up-Mix Stereo to Surround Based on MPEG Surround Technology," presented at the *122nd Convention of the Audio Engineering Society* (2007 May), convention paper 6991.

[27] F. Rumsey, "Controlled Subjective Assessments of Two-to-Five-Channel Surround Sound Processing Algorithms," *J. Audio Eng. Soc.*, vol. 47, pp. 563–582 (1999 Jul./Aug.).

[28] ITU-R, "Report ITU-R BS.1116-2 Methods for the Subjective Assessment of Small Impairments in Audio Systems," *International Telecommunications Union* (2014).

[29] H. Lee and C. Gribben, "Effect of Vertical Microphone Layer Spacing for a 3D Microphone Array" *J. Audio Eng. Soc.*, vol. 62, pp. 870–884 (2014 Dec.). http://dx.doi.org/10.17743/jaes.2014.0045

[30] M. Barron and A. Marshall, "Spatial Impression due to Early Lateral Reflections in Concert Halls: the Derivation of a Physical Measure," *J. Sound Vib.*, vol. 77, pp. 211–232 (1981). http://dx.doi.org/10.1016/S0022-460X(81)80020-X

APPENDIX

The direct to reverberant (D/R) energy ratios presented in Table 1 are calculated using the following equation.

$$D/R\ energy\ ratio = \frac{\int_{5ms}^{\infty} p^2(t)dt}{\int_{0}^{5ms} p^2(t)dt}[dB], \tag{1}$$

where $p(t)$ is the instantaneous sound pressure of room impulse response. The division value of 5 ms was based on [4].

Interchannel cross-correlation coefficient (ICCC) is defined as the maximum absolute value of the normalized cross-correlation function (NCF), which is expressed as below.

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

where $x_1(t)$ and $x_2(t)$ are the instantaneous sound pressures of room impulse responses, t1 and t2 are the lower and upper boundaries of time segments, and τ is the time lag. The measurement results shown in Table 2 were obtained for two time segments of room impulse responses separately: t1 = 0ms to t2 = 80ms (early sound) and t1 = 80ms to t2 = 750ms (late sound). The division values are based on [4]. The lag (τ) limit was the length of each segment.

THE AUTHOR

Hyunkook Lee

Hyunkook Lee received a B.Mus. degree in music and sound recording (Tonmeister) from the University of Surrey, Guildford, UK, in 2002, and his Ph.D. degree in audio engineering and psychoacoustics from the Institute of Sound Recording (IoSR) at the same University in 2006. From 2006 to 2010, Dr. Lee was Senior Research Engineer in audio R&D at LG Electronics, South Korea. Currently, he is Senior Lecturer in music technology and the leader of the Applied Psychoacoustics Lab (APL) at the University of Huddersfield, UK. Dr. Lee has also been a freelance recording engineer since 2002. His current research includes spatial audio psychoacoustics, 3D sound recording and image rendering techniques, and interactive virtual acoustics. He is an active member of the Audio Engineering Society and a fellow of the Higher Education Academy, UK.