Lee, Hyunkook, Gribben, Christopher and Wallis, Rory

Psychoacoustic Considerations in Surround Sound with Height

Original Citation

Lee, Hyunkook, Gribben, Christopher and Wallis, Rory (2014) Psychoacoustic Considerations in Surround Sound with Height. In: 28th Tonmeistertagung: tmt 28, 20th-23rd November 2014, Cologne, Germany.

This version is available at http://eprints.hud.ac.uk/23151/

Psychoacoustic Considerations in Surround Sound with Height
(Psychoakustische Betrachtungen zu Surround Sound mit Höhenabbildung)

H. Lee, C. Gribben and R. Wallis
Applied Psychoacoustics Laboratory (APL), University of Huddersfield, Huddersfield, HD1 3DH, UK

Abstract

This paper presents recent research findings in the psychoacoustics of 3D multichannel sound recording and rendering. The addition of height channels in new reproduction formats such as Auro-3D, Dolby Atmos and 22.2, etc. enhances the perceived spatial impression in reproduction. To achieve optimal acoustic recording and signal processing for such formats, it is first important to understand the fundamental principles of how we perceive sounds reproduced from vertically oriented stereophonic loudspeakers. Recent studies by the authors in this field provide insights into how such principles can be applied for practical 3D recording and upmixing. Topics that are discussed in this paper include the interchannel level and time difference relationships in terms of vertically induced interchannel crosstalk, the effectiveness of the precedence effect in the vertical plane, the aspect of tonal coloration resulting from vertical stereophonic reproduction, the effect of vertical microphone spacing on envelopment, the effect of interchannel decorrelation, and the use of spectral cues for extending vertical image spread.

1. Introduction

Height channel loudspeakers used in new 3D multichannel audio formats such as Auro-3D [1] and Dolby Atmos [2] add the height dimension to the width and depth dimensions existing in the conventional surround formats. The added height channels are naturally expected to enhance perceived spatial impression, however the optimal ways to achieve the goal have not yet been fully understood. There are basically two ways of mapping signals to the height channels; one could simply route a discrete source or ambient signal to one of the height loudspeakers, or attempt to create a vertical phantom image between main and height loudspeakers. The former is a case of the localisation of an elevated monophonic source. For more flexible panning of the perceived image and spatial rendering in 3D recording and mixing, one would adopt the latter scenario where two coherent signals are fed into both of the main and height loudspeakers and processed, in order to produce the desired perceptual effects. However, there is a question of how effective the conventional interchannel signal manipulation methods, such as amplitude panning, time panning and decorrelation, would be perceived for the processing of vertically oriented signals. This is fundamentally because signals radiated by the main and height loudspeakers would cause no or minimised interaural time and level differences. Furthermore, added height channels might even cause unpleasant colouration of sound due to comb-filtering that would occur if the main and height signals arriving at the ear had time or phase differences – research suggests that a reflection arriving from the same direction as the direct sound causes a more negative tone colouration than that arriving from other directions [3]. Therefore, in order to develop perceptually optimised methods for 3D sound recording and rendering, it would be first important to understand the psychoacoustic principles of how we perceive auditory image in vertical stereophony. Although there has been much research conducted on the vertical localisation of monophonic sound [4, 5, 6, 7, 8. 9], the perceptual mechanism of vertically oriented stereophonic sound has been relatively under-researched. Most widely referred literature includes the work by Barber [10], which reported that the conventional amplitude panning did not work effectively between loudspeakers placed in the median plane, and the VBAP by Pulkki [11], which is a method for 3D amplitude panning using three energy weighted vectors. Recently, the authors of this paper and colleague researchers have carried out various psychoacoustic experiments in the context of vertical stereophony for 3D format and reported the results. This paper reviews some of these studies and discusses the practical implications of the findings. The topics covered in this paper are as follows:

- Relationship between interchannel level and time differences for localisation and audibility of vertical stereo
- Effect of vertical interchannel decorrelation
- Effect of vertical microphone layer spacing for a 3D microphone array
- Rendering of vertical image spread based on perceptual band allocation (PBA)

2. Relationship between Interchannel Level and Time Differences for Vertical Stereo

In horizontal stereophony, phantom image localisation relies on the summing localisation process [6]; within a certain limit of interchannel time difference (ICTD) (e.g. < 1ms), an interchannel level difference (ICLD) or ICTD, or the combination of the two applied between two loudspeaker signals is translated into some combination of interaural level difference (ILD) and interaural time difference (ITD) when the loudspeaker signals are summed at each ear with acoustic crosstalk. Beyond the ICTD limit, however, the precedence effect operates and the localisation of phantom image is dominated by the ICTD; the image is perceived at the direction of the earlier loudspeaker and a trade-off between ICLD and ICTD is not possible any more. In the case of vertical stereophony, on the other hand, these psychoacoustic rules would not apply in the same way since the role of binaural cues is minimised.

2.1. Recent Research

Lee [12] subjectively measured the minimum ICLD between main and height loudspeaker signals that is required for the perceived image to be localised fully at the position of the image produced by the main loudspeaker (referred to as 'localised threshold'). The task for subjects was to reduce the level of the height loudspeaker until the location of the resultant sound image was perceived to match that of the main loudspeaker, which was the reference. Anechoically recorded cello and bongo performance excerpts were used as the sound sources, and eight critical listeners participated in the listening tests. The main and height loudspeakers were vertically arranged, with the former and latter elevated at 0° and 30° from the listening position. Nine different time delay values ranging from 0 to 50ms were applied to the height loudspeaker. A similar test was repeated to measure the minimum ICLD for any perceptual effects of the height channel signal to be completely masked by the sound of the main channel only (referred to as 'masked threshold'). The results showed that with the ICTDs up to 5ms, the amount of level reduction required of the height channel signal was relatively constant for both localisation and masking, and that the ICTD had no statistically significant effect on the subjects' judgments. The average localised threshold in this ICTD range was between -6 and -7dB while the average masked threshold was between -9 and -10dB, as can be observed in Fig. 1. These results suggest that the precedence effect does not operate between vertically arranged loudspeakers in the median plane. This seems to disagree with Litovesky et al. [13] who reported that the precedence effect still operated in the median plane. However, their experiment was conducted with front, overhead and rear loudspeakers placed in the median plane, whereas Lee's experiment was with loudspeakers vertically elevated at 0° and 30° from the listener's ear level. Furthermore, their research was focused on the question of whether the localisation of a perceived image is 'dominated' by the leading sound rather than whether the image is fully localised at the direction of the leading sound.

Stenzel et al. [14] conducted experiments similar to Lee's with a diagonally oriented loudspeaker pair instead of a vertically oriented one, using anechoic speech, cello and conga sources. Similarly to Lee's findings, they found that neither localisation nor masking depended on the ICTD up to a certain limit (0 to 10ms). The average level reduction applied to the height channel signal for localisation was 6dB, which is similar to Lee's result. This means that the precedence effect did not operate even with a diagonally arranged loudspeaker pair, where some interaural differences are presented. On the other hand, the threshold for masking was -16dB, which is substantially greater than that Lee's 9 to 10dB, but close to masked thresholds of a horizontally arriving reflection reported in the literature [3, 15, 16]. This seems to be associated with the presence of the interaural differences and their perceptual effects.

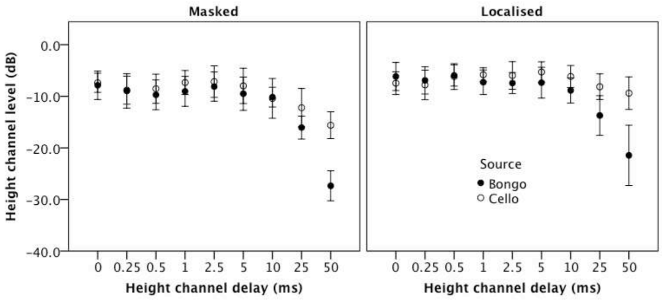

Fig. 1: Masked (left panel) and localised (right panel) thresholds of vertical interchannel crosstalk as a function of height channel delay time.

Wallis and Lee [17] conducted a more controlled experiment on the vertical localisation of stereophonic signals in an anechoic chamber using octave-band noise with centre frequencies from 125 to 8000Hz as well as a broadband noise. The loudspeaker arrangement was identical to Lee [12]'s, but an acoustically transparent curtain was placed directly in front of the loudspeakers in order to obscure the test setup from the listeners. They conducted two experiments. The first experiment examined the effect of ICTD alone on vertical stereophonic localisation. For this experiment, stimulus presentation was either from an individual loudspeaker (monophonic) or from both loudspeakers together (stereophonic). For the stereophonic condition, the height loudspeaker was delayed with respect to the main loudspeaker by 0, 0.5, 1, 5 or 10ms. Fig. 2 shows the results for the lower (main) and upper (height) loudspeaker conditions as well as for the 0 and 10ms ICTD conditions. As can be seen, the effect of ICTD on the localisation of octave band stimuli in the median plane is highly erratic. Comparing between the 0ms and 10ms conditions, there is no conventional panning effect observed for any of the frequency bands; a longer height channel delay does not cause the image to be shifted more towards the main loudspeaker. Rather, a more interesting tendency is that the octave band results for stereophonic presentation look somewhat similar to those for monophonic presentation. The localisation appears to be primarily governed by the so-called 'pitch-height' effect [4, 5], which suggests that a higher frequency tends to be localised at a higher position, rather than the presentation method. Past research suggests that the relationship between the pitch-height effect and loudspeaker height depends on the bandwidth of the signal; a pure tone or narrow band signal are localised by the pitch-height effect regardless of the loudspeaker's actual height, whereas a wide band signal containing frequencies above around 7kHz are localised accurately at the height of its presenting loudspeaker.

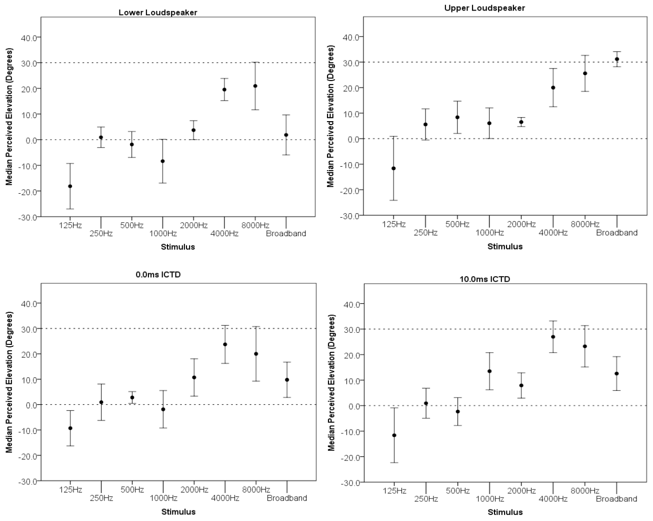

Fig. 2: Perceived elevation of octave and broadband stimuli for lower and upper loudspeaker presentations (top left and top right) as well as the 0ms (bottom left) and 10ms (bottom right) stereophonic conditions.

The second experiment by Wallis and Lee [17] measured localised thresholds for octave bands of noise using the same method used for the localised threshold measurement. This experiment therefore considered the relationship between vertical ICTD and ICLD. The experimental setup and the ICTD values tested were identical to the first experiment. It was found that the effect of ICTD was statistically not significant. As can be seen in the results plotted in Fig. 3, the localisation thresholds averaged for all ICTDs were not uniform across different frequency bands. The thresholds for the low frequencies were found to be fairly high (-5.3dB at 125Hz, -3.03dB at 250Hz). This decreased dramatically to between -9 and -10.5 dB when the frequency increased beyond 1000 Hz. Of interest in the result was the agreement in the threshold for both the 500 and 4000 Hz octave bands and the results obtained for musical sources by Lee [12]. As both 500 and 4000 Hz are related to frontal perception according to Blauert's directional bands [6], it might be suggested that these frequencies had some perceptual dominance in the localisation thresholds obtained in Lee's study.

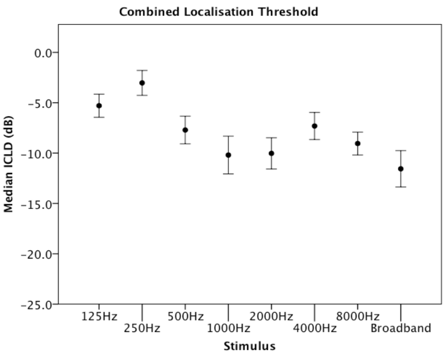

Fig. 3: Median localised thresholds for both octave band and broadband stimuli.

2.2. Discussion

The findings from the above studies have useful implications for several practical 3D audio applications. One way of using the height channels for 3D music production would be a vertical extension of sound image. One might desire to achieve this while still locating the source image at the main loudspeaker height. This would be particularly true for classical recordings made using a 3D microphone array in an acoustic environment, where the height channels are most likely to be used for presenting extra ambience information rather than panning source images vertically (except for vertically long sources such as the organ or a large choir on layered stands). Then it would be aimed to extend the impression of listener envelopment (LEV) vertically without shifting the source image position upwards from the main loudspeaker height. The localised threshold founded in [12] could be used for the configuration of vertically arranged microphones – the upper microphone would need to be at least 6 to 7dB attenuated compared to the lower microphone, so that the direct sound picked up by the upper microphone (i.e. vertical interchannel crosstalk) would not affect the localisation of phantom image – this can be done by adopting a directional polar pattern and angling the microphone appropriately. For instance, a vertically coincident XY cardioid pair can be configured at the subtended angle of at least 90°, with one microphone facing towards the source and the other facing upwards. A spaced omni-directional main and height microphone pair approach would not ensure the phantom image to be localised at the main loudspeaker layer since an ICTD is not effective in vertical panning as discussed above. Furthermore, the lack of ICLD in the omni microphone signals would also give rise to potentially unpleasant colouration of sound when the signals are summed monaurally at the ear. However, the 6 to 7dB level attenuation of vertical crosstalk would still not be sufficient to suppress perceptual effects that could be potentially perceived between vertically arranged loudspeakers. In the present authors' pilot experiment, such effects as vertical source image spread and tone colouration were perceived at the localised threshold, depending on the ICTD applied, and it was found that they could be positive factors depending on the source material.

However, if these effects were not desired by the sound engineer, the height channel level would need to be further reduced to below the masked threshold, which is -9 to -10dB [12]. Furthermore, Stenzel et al.'s findings suggest that if perceptual effects between diagonal pairs of loudspeaker (e.g. front left and front right height) is of concern, then the levels of the height channels should be reduced at least by 16dB. In terms of microphone technique, a maximum suppression of vertical crosstalk can be achieved by using a figure-of-eight microphone with its null-point facing towards the source.

The localised thresholds for octave bands measured by [17] might be useful for the vertical rendering of source image. As the localisation threshold was found to be dependent on frequency, controlling ICLD for each individual octave band is likely to be more effective than applying a "blanket" ICLD when trying to create spatial or tonal effects, while maintaining the localisation of a sound source at the position of the lower loudspeaker.

3. Vertical Interchannel Decorrelation

Interchannel decorrelation is a popular method of generating a wider auditory image width or spread between two channels, by altering the phase content of a monophonic source. Following decorrelation, the resultant waveforms of the processed signals must be significantly different from one another (in order for spread to be perceived) though sonically similar, and also demonstrate a low value of interchannel cross-correlation (ICCC) (where 0 is fully decorrelated). However, little research has been carried out in the area of vertical decorrelation.

3.1. Recent Research

Gribben and Lee [18] conducted experiments to fundamentally investigate interchannel decorrelation both horizontally and vertically, comparing its effect between the two domains. To date various decorrelation techniques have been proposed, mainly surrounding the use of comb and all-pass filters to alter a monophonic signal. The so-called 'Lauridsen' decorrelator [19] was used in this experiment for its simplicity and flexibility for controlling the degree of signal correlation. Listening tests were conducted in an anechoic chamber using two loudspeakers. For the horizontal stereo setup, the loudspeakers were at +/- 30° azimuth to the listener; and during the vertical test, upper and lower loudspeakers were placed in the median plane, with the upper loudspeaker at a vertical angle of 30° to the listener. For stimuli, a monophonic pink noise sample was filtered into three frequency-bands, with each band consisting of three octaves: low (octave bands with centre frequencies of 64Hz, 125Hz and 250Hz), mid (500Hz, 1kHz and 2kHz) and high (4kHz, 8kHz and 16kHz). These samples were processed varying the parameter values for time-delay (1ms, 5ms, 10ms and 20ms) and the decorrelation factor (0-1, where 0 is no decorrelation and 1 is fully decorrelated).

Observing the horizontal decorrelation results averaged for all time-delays, Fig. 4 shows a strong linear relationship between the decorrelation factor and perceived auditory source width for every frequency-band. These results indicate that the value of the decorrelation factor effectively increases the width of the signal, relative to a reference which is perfectly correlated, particularly for the mid and high frequency bands. Low frequency results show an overlap of error-bars for the decorrelation factors of 0.6-1.0, representing no significant difference for these values. Analysing the time-delays individually showed a longer time-delay was more effective for low frequency content. The effect of time-delay was largely insignificant for the mid and high frequency-bands, although a slight improved decorrelation was also demonstrated as the time-delay increased for these frequency-bands. The vertical results shown in Fig. 4 also demonstrate some linearity between the decorrelation factor and perceived vertical spread, although to a lesser extent than the horizontal results. It appears that, in the case of vertical decorrelation, the low frequency band is most effective, however, significant overlap can be seen between all results, indicating that there is no significant difference amongst close decorrelation factors. Many results for the high frequency band showed an insignificant difference from the 0 reference, indicating that decorrelation had little effect in terms of spread increase for these frequencies.

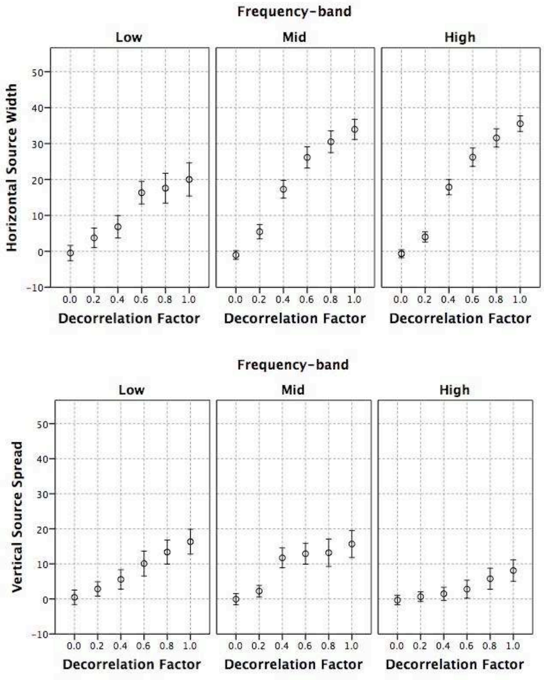

Fig. 4: Horizontal (upper panel) and vertical (lower panel) decorrelation Results, mean values and associated 95% confidence intervals.

Comparing the horizontal and vertical results in Fig. 4, it can be seen that the decorrelation is effective and perceptible, to a certain extent, in both domains. For horizontal decorrelation, it seems that the higher the frequency, the more effective the process is; whereas, vertical decorrelation appears to be most effective for lower frequencies. In general, the listeners used a greater scale when grading the horizontal stimuli, which may demonstrate that the audible difference between stimuli is more prominent and easier to perceive in this domain. All results show some linear increase between decorrelation factor and perceived spread, which is the desired trend, though there is significant overlap for the vertical results, signifying that decorrelation is only truly perceptible when using higher decorrelation factors.

Given that the test focused on the assessment of the relative spread between decorrelation factors, the absolute spread of the initial reference prior to decorrelation was not gauged; therefore, further investigation into the degree of spread increase from the reference is required to determine the overall effect of decorrelation. This is of particular interest for the vertical high frequency results, as there is seemingly such little difference of perceived spread between the 1.0 decorrelation factor and reference, whereas the high frequency band was noticeably the most effective for horizontal decorrelation.

3.2. Discussion

Filtering based decorrelation methods are popular for mono to stereo or stereo to surround upmixing. Such methods might also be used for the rendering of vertical stereo images. However, one potential concern with using the decorrelation method in the vertical domain is that when two vertically arriving signals with different phase relationships are summed at the ear, an audible and possibly undesirable colouration of sound might be caused due to changes in phase-amplitudes depending on frequencies (i.e. the comb-filter effect). Tone colouration occurring between vertically arriving sounds would typically be more problematic in terms of perceived sound quality than that between horizontally arriving sounds [3]; in the case of the conventional 2-channel horizontal stereo, a sound reproduced by the contralateral loudspeaker is attenuated at high frequencies due to a diffraction effect of the listener's head, thus potentially less problematic comb-filtering. This problem could be made worse when multiple pairs of vertically arranged loudspeakers are used at different azimuths. From the fact that a comb-filtering effect is typically more audible at high frequencies, and also from Gribben and Lee's results showing that the decorrelation of low frequencies was more effective for controlling vertical image spread than that of high frequencies, it might be more reasonable to apply decorrelation only to low frequencies than to all frequencies in 3D upmixing or image rendering applications using vertical decorrelation.

4. Vertical Microphone Layer Spacing for a 3D Microphone Array

In horizontal stereophony, it is well known that a more spaced microphone pair would produce a greater spatial impression owing to the decrease in interchannel correlation [19, 20]. However, as discussed in the previous section, interchannel decorrelation in the vertical domain would not be as effective as that in the horizontal domain. Therefore, the effect of spacing between main and height channel microphones on perceived spatial impression and sound quality could be questioned.

4.1. Recent Research

Recently this has been investigated by Lee and Gribben [21] in the context of a 3D main microphone array. Recordings were made in a reverberant concert hall using a 5-channel main array augmented with four upward-facing cardioid microphones for height channels. The height microphones were placed directly above the front left, front right, rear left and rear right microphones of the main array, at four different vertical spacings of 0m, 0.5m, 1m and 1.5m (Fig. 5). Sound sources tested were trumpet solo, acoustic guitar solo, percussion quartet and string quartet. Multichannel room impulse responses were also acquired for objective measurements of various channel and binaural signal relationships, and the spectral energy influence of the height layer. Listening tests were conducted using a 9-channel loudspeaker setup based on an Auro-3D 9-channel format.

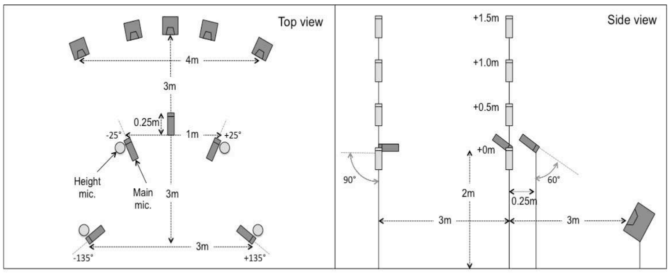

Fig. 5: Microphone setup for comparing different spacings of height microphone layer.

It was mainly found that the layer spacing had a minor effect on both the perceived spatial impression and overall quality preference. The layer spacings of 0.5m, 1m and 1.5m had no statistically significant differences for all sources. As can be seen in Fig. 6, the 0m layer was found to produce slightly greater spatial impression than the spaced layers for the guitar and percussion sources, but there was no significant difference between the 0m and the more spaced layers for the trumpet and strings. These results seem to be associated with the perceptual effect of vertical interchannel crosstalk (direct sounds in the height channel signals). The ICLDs between the main and 0m height layer signals were 7.6dB. According to Lee [12]'s localised (-6 to -7dB) and masked thresholds (-9 to -10dB), this ICLD value is large enough to locate the sound image at the main loudspeaker layer but not enough to completely suppress the potential effects of crosstalk. However, the ICLD values for all the spaced pairs were larger than 10dB, and therefore no or little audible crosstalk effects would have been produced. The energies of the ambient parts of the height channel room impulse responses for all spacings were almost constant. Interchannel cross-correlation coefficients (ICCCs) for the main and height channel signals with different spacings were measured in octave bands for the main and height channel impulse responses with different spacings. It was found that the ICCCs for the ambient parts decreased linearly as the spacing increased. This does not seem to agree with the perceived spatial impression results. It also seems contradictory to Gribben and Lee [18]'s results shown above; low frequency decorrelation was more effective for increasing vertical image spread than mid or high frequency decorrelation. However, it is worth noting that Gribben and Lee's experiment only used a single pair of loudspeakers in the median plane, whereas the current experiment used a 9-channel loudspeaker format. Binaural signal energies of the ambient sounds measured in the listening position were found to be constant. Furthermore, interaural cross-correlation coefficients for reverberant sounds ($\mathrm{IACC}_{\mathrm{late}}$) measured at the listening position hardly varied with different layer spacings. From the above discussions, it can be suggested that among the spaced microphone layers the binaural ambience energy for each layer and the horizontally perceived LEV (i.e. $\mathrm{IACC}_{\mathrm{late}}$) played a more dominant role in the perception of overall spatial impression, whereas the level of vertical crosstalk was the main reason for the slight 0m layer dominance over the spaced layers.

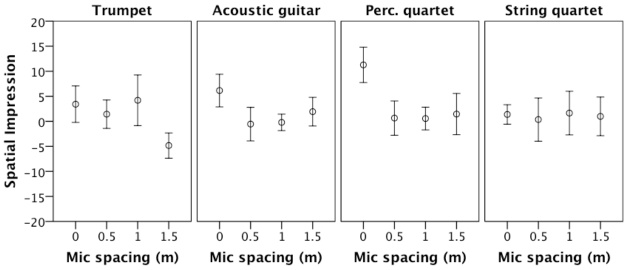

Fig. 6: Microphone spacing vs. spatial impression for each source: Mean values and associated 95% confidence intervals.

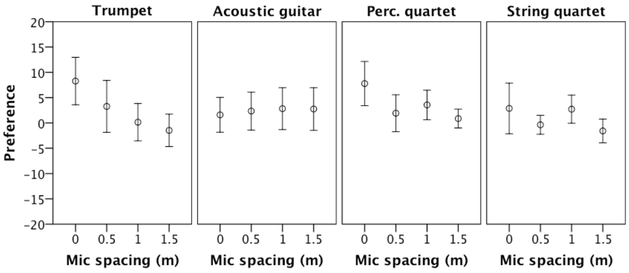

Fig. 7: Microphone spacing vs. preference for all sources: Mean values and associated 95% confidence intervals.

The preference results also showed no significant difference between spaced layers, but the 0m coincident layer was rated slightly higher than the spaced layers depending on sound source (Fig. 7). The main preference attributes commented by the listeners were tonal quality as well as spatial quality, and this can be discussed in relation to the delta spectrum measurements shown in Fig. 8. The frequency responses of the direct and ambient sound components of the left ear input signal for the main layer was subtracted from those for the main layer combined with each height layer, respectively. It can be seen that the addition of the coincident microphone layer to the main layer had a positive and almost flat spectral influence on the ear signal spectrum, whereas that of a spaced layer produced comb-filter effects in the resulting spectrum.

Scuda et al. [22] compared the performances of two practical 3D microphone arrays, one with vertically coincident microphone layers and the other with vertically spaced layers, in a 3D reproduction environment using castanets, cello and speech sources. The attributes tested were localisation performances in both horizontal and vertical angular displacements, perceived width and perceived spaciousness. They found that in overall the vertically coincident array performed better than or similar to the spaced array in all of the attributes they tested, depending on the sound source. Although this experiment tested two practical techniques with two different design concepts, thus not strictly controlling experimental variables other than the spacing of the vertical layer, the results seem to suggest that a vertically coincident microphone configuration can perform better than or at least similar in result to a vertically spaced configuration.

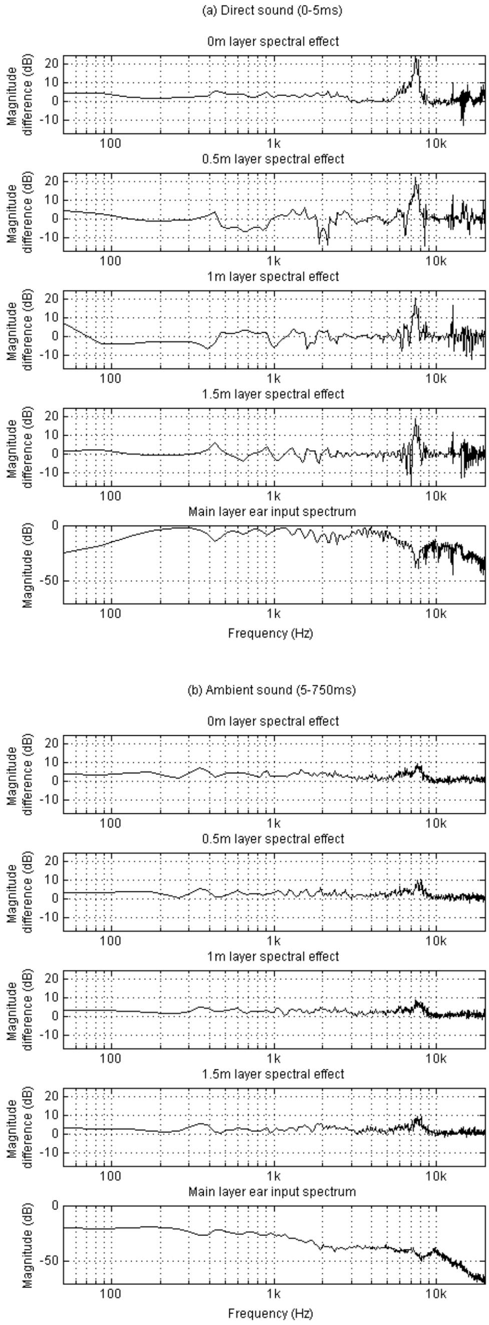

Fig. 8: Delta spectrum between the left ear impulse response of the main and height layers and that of the main layer only [21]; (a) direct sounds (0ms…5ms) (b) ambient sounds (5ms…750ms); the bottom panels show the magnitude frequency response of the ear input signal for the main layer.

4.2. Discussion

From the above findings, it might be suggested that in practical recording situations, where vertical interchannel crosstalk is inevitably present due to the desired configuration of height microphone (e.g. upward facing cardioid), a vertically coincident 3D main microphone array could be beneficial compared to a vertically spaced array, since it would not suffer from the comb-filtering of source signal at the ear. The characteristics of ambient signal picked up by the coincident height microphone would be similar to those by a more spaced height microphone, at least within the range of the spacings tested in Lee and Gribben [21]'s study (0 to 1.5m). The potential benefit of a larger channel separation (i.e. ICCC) achieved by applying a larger vertical spacing between microphones also seems minimal within such a range. However, this needs to be confirmed for a wider range of microphone spacing as well as that of source type. The coincident nature of main and height channel signals will also be useful for 3D to 2D downmix applications.

In recording situations where the suppression of interchannel crosstalk in height channels is the priority over the polar pattern and angle of height microphones (i.e. a total masking of the perceptual effects of height channel crosstalk), the following microphone configurations could be adopted. In terms of a vertically coincident setup, a maximum rejection of direct or source sound in the height microphone could be achieved by using a so-called 'back-to-back' cardioid configuration (i.e. 180° subtended angle between the microphones). Alternatively, a figure-of-eight height microphone could be adopted so that its null-point faces towards the source, although in this case the rear lobe of the microphone might pick up unwanted floor reflections or audience noise. If a vertically spaced array is desired by the sound engineer to achieve greater channel separation (e.g. lower ICCC), it is suggested that the omni-directional microphone pair, which lacks ICLD, would not be ideal from both localisation accuracy and tonal quality points of view. As discussed in Section 1, the precedence effect does not fully operate in a vertical stereophony and therefore an ICTD between the omni-directional microphones would not be useful for localisation; from Wallis and Lee [17]'s octave-band localised threshold measurements, it can be anticipated that the perceived phantom image would be localised somewhere between the main and height loudspeakers and the image position might fluctuate up and down over time depending on the time-varying spectrum of the signal. The lack of vertical ICLD in the presence of vertical ICTD also means that there would be strong comb-filtering when the main and height channel signals are summed at the listener's ear, which might be perceptually unpleasant. Note that the issue of comb-filtering is not serious for diffused ambience signals as can be seen in Fig. 8(b). This implies that there would be a greater degree of freedom to choose the microphone's polar patterns and angles beyond the critical distance of a recording venue when the dedicated ambience microphone array that has been separated from the main array is used. Investigations into the optimal 3D microphone configurations, and the effect of ICCC on spatial perception for diffused reverberant sounds, are currently being carried out by the first author of this paper.

5. Vertical Image Spreading based on Perceptual Band Allocation (PBA)

Past research into the effects of spectral cues on vertical localisation [4] generally suggest that a sound source with a higher frequency tends to be localised at a higher position than that with a lower frequency. This phenomenon is often referred to as the 'pitch-height' effect. It was confirmed by Roffler and Butler [5], Morimoto et al. [7] and Cabrera and Tiley [8] that this effect is valid with band-passed noise signals as well as pure tones, and that the loudspeaker height dependency of vertical localisation was related to the bandwidth of the signal. Broadband signals or signals containing frequencies above around 7kHz are localised accurately in the vertical plane, whereas signals with a lack of high frequencies tend to be localised poorly, regardless of the loudspeaker's physical position [5]. It was further found by Ferguson and Cabrera [9] that low-pass and high-pass filtered noise signals simultaneously presented by two vertically arranged loudspeakers also produced the pitch-height perception; the low-passed band was perceived at the listener's eye-level regardless of the loudspeaker's physical height, whereas the high-passed one presented from an elevated loudspeaker was localised around the height of the loudspeaker. Furthermore, it was shown in [8] that the vertical image spread of an individual octave band noise could be as large as that of a broadband noise.

5.1. Recent Research

Lee [23] investigated whether the pitch-height effect described above could be exploited for the creation of the auditory sensation of 3D LEV (i.e. engulfment) in the context of 2D to 3D ambience upmixing. Recordings were made in a reverberant concert hall with a 3-channel frontal microphone array and a 4-channel ambience array (2m x 2m Hamasaki-Square [20]) which was augmented with four extra height microphones as can be seen in Fig. 9. The augmented Hamasaki-Square was a novel approach as no research had been reported on the optimal method for capturing vertical stereophonic ambience before this study; as mentioned in Section 3.2, this research is currently ongoing. The 3D ambience array was placed beyond the critical distance of the recording venue to ensure the D/R energy ratio was below 1, and a 1m spacing was applied between the main and height layer microphones. As shown in Fig. 8(b), applying a vertical spacing between microphones in a diffused field would not be of critical concern in terms of the comb-filtering of direct sound, although it could still affect the tonal colour of the reproduced ambient sound. The 1m spacing was generally found to provide a more pleasant tonal colour (but not necessarily a greater spatial impression) than a 0m spacing from a pilot experiment. The vertical ICCCs for the left main and left height ambience microphones were 0.25, 0.1 and 0.1 for low (125Hz and 250Hz) and mid (500Hz, 1kHz and 2kHz) and high (4kHz and 8kHz) octave frequency bands, respectively, which seems to be reasonably low for low frequencies. A larger spacing, e.g. 2m, would lower the ICCC, but its effect might not be great according to Gribben and Lee [18]'s findings described in Section 2. Especially, a small difference in high frequency decorrelation is very difficult to resolve vertically.

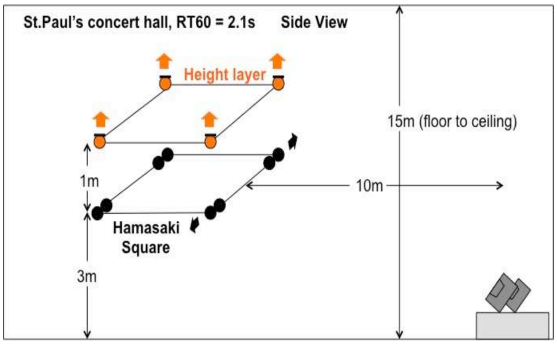

Fig. 9: Augmented Hamasaki-Square for capturing ambience for 3D reproduction.

For 2D to 3D upmixing, the original Hamasaki-Square 4-channel ambience signals were low- and high-pass filtered at three different crossover frequencies: 0.5k, 1k and 4kHz. The low-passed signals were mapped to the main loudspeakers, with the high-passed ones to the height loudspeakers, in a 9-channel Auro-3D-inspired configuration. This mapping method for vertical upmixing was referred to as the 2-band 'perceptual band allocation' (PBA). Listening tests were conducted to rate the following mixing conditions in terms of the magnitude of perceived 3D LEV: original 9-channel 3D mix, original 5-channel 2D mix, 7-channel 3D mix with the Hamasaki-Square ambience mapped to the height channels instead of the main channels, and PBA-upmixed versions of the 2D mix at the three crossover frequencies. 3-channel 1D mix (front three microphones only) was also included as a low anchor. The playback levels of all 3D stimuli were aligned as reference to the original 9-channel mix at the listening position. Sound sources used were conga solo, trumpet solo and string quartet.

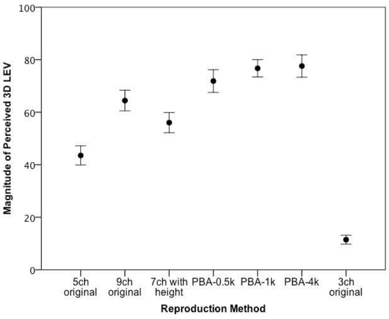

Fig. 10: Mean values and the associated 95% confidence intervals of the 3D LEV data for all sources.

Overall test results showed that the PBA methods generally produced a similar or greater magnitude of perceived 3D LEV compared to an original 9-channel 3D recording (Fig. 10). However, individual results for each sound source showed a significant interaction between the reproduction method and sound source (Fig. 11). The strings had no significant differences between either the 9-channel mix and any of the PBA upmixes, or among all of the PBA upmixes. For the conga, all the PBAs were rated significantly higher than the 9-channel mix, whilst there was no significant difference between any of the PBAs. For the trumpet, on the other hand, the PBA rating increased linearly and significantly as the cutoff frequency increased, and only the 4kHz-PBA was significantly higher than the 9-channel original.

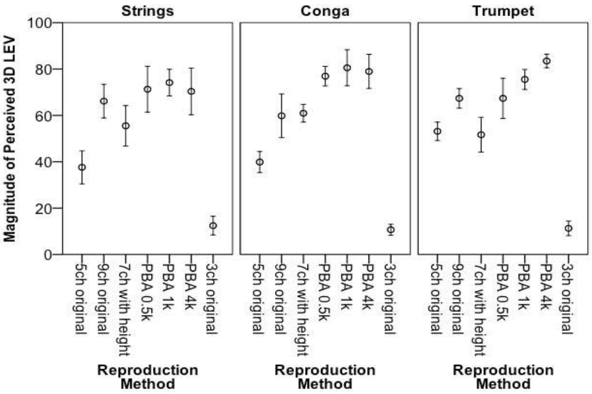

Fig. 11: Mean values and the associated 95% confidence intervals of the 3D LEV data for each source.

5.2. Discussion

The above results were somewhat unexpected in that the original 3D recording, which was initially considered as the highest quality reference, was graded lower than the upmixed versions. This might be associated with a potential limitation of the ambience recording technique used in this study or the acoustic condition of the venue. However, it also suggests that the PBA method could produce a 3D LEV impression that is comparable to an original 3D recording. The reason for the 9-channel sound to produce less LEV than PBA for the conga was mainly related to the boost of certain low-mid frequencies, which made the localised generally lower by the pitch-height effect. Moreover, the conga source did not have much high frequency energy. One of the advantages of the PBA method is that there is no overlapping of frequencies at the ear, thus no comb-filtering; vertical image spread is rendered while maintaining the spectrum of the original ambience. Informal preference rating tests showed that all the PBAs were preferred to the 9-channel mix, especially for the conga, mainly due to the tonal quality.

The PBA method could be extended for multiple frequency bands to give a finer control of vertical image spread. Results from past research on the pitch-height effect [6, 8] imply that each frequency band of a signal would have its own perceptual position in the vertical plane, depending on the height of the loudspeaker that presents it. From this, it is hypothesised that if the perceptual position of each sub-band for each of the main and height loudspeakers in a 3D format were measured, a flexible mapping between band and loudspeaker would be possible for the control of vertical image width. This is currently being investigated by the first author of this paper.

6. Conclusions

This paper reviewed and discussed some of the recent psychoacoustic studies conducted on 3D sound recording and signal processing. From this, the following conclusions can be drawn.

- Psychoacoustic principles for 3D multichannel audio are different from those for horizontal stereo. This should be considered in the development of new 3D audio technologies, such as 3D microphone arrays, and vertical image rendering or upmixing algorithms.
- In 3D microphone array design, height microphones should be configured with the vertical interchannel level and time difference relationship in mind (e.g. level and delay of vertical crosstalk) - Care must be taken not only on spatial impression, but also on vertical localisation and tonal colouration.
- The precedence effect does not operate vertically (at least not with the 30° loudspeaker elevation), but the pitch-height effect tends to govern the localisation of a delay-based vertical stereo image.
- Interchannel decorrelation for vertical image spreading is not as effective as that for horizontal spreading. It seems that the vertical effect is greater for low and mid frequencies than for high frequencies.
- The pitch-height effect can be exploited for the extension of vertical image spread (e.g. Perceptual Band Allocation (PBA) for ambience upmixing)

The present authors are currently working on the elicitation of 3D audio attributes, the development of an objective model for 3D audio quality evaluation, and the development of a 3D virtual acoustic renderer for music production. Some of the results from these studies will be presented at Tonmeistertagung 29 in 2016.

7. Acknowledgement

This work was supported by the Engineering and Physical Sciences Research Council (EPSRC), UK, Grant Ref. EP/L019906/1.

8. References

[1] B. V. Daele and W. V. Baelen, Productions in Auro-3D, URL: http://www.auro-3d.com/professional/technical-docs/. 2012.

[2] Dolby, Authentic Cinema Sound by Dolby Atmos, URL: http://www.dolby.com/gb/en/consumer/technology/movie/dolby-atmos-details.html.2014

[3] M. Barron, “The Subjective Effects of First Reflections in Concert Halls – The Need for Lateral Reflections,” *J. Sound Vib.*, vol. 15, pp. 475–494, 1971.

[4] C. C. Pratt. “The Spatial Character of High and Low Tones,” *J. Exp. Psy.*, vol. 13(3), pp. 278–285. 1930.

[5] S. K. Roffler and R. A. Buttler, “Factors that influence the localisation of sound in the vertical plane,” *J. Acoust. Soc. Am.*, vol. 43 (6), pp. 1255-1259 (1968).

[6] J. Blauert, *Spatial Hearing*, rev. ed. (MIT Press, Cambridge, MA, 1997).

[7] M. Morimoto, M. Yairi, K. Iida, and M. Itoh, “The Role of Low Frequency Components in Median Plane Lo- calization,” *Acoust. Sci. Technol.,* vol. 24, pp. 76–82 (2003).

[8] D. Cabrera and S. Tilley. “Vertical Localization and Image Size Effects in Loudspeaker Reproduction,” In *Audio Engineering Society 24th International Conference: Multichannel Audio*, *The New Reality*. 2003.

[9] S. Ferguson and D. Cabrera. “Vertical Localization of Sound from Multiway Loudspeakers,” *J. Audio Eng. Soc.,* 53(3), 163–173. 2005.

[10] J. L. Barbour, “Elevation Perception: Phantom Images in the Vertical Hemi-Sphere,” In *Audio Engineering Society 24th International Confererence: Multichannel Audio*. *The new Reality.* 2003.

[11] V. Pulkki, “Virtual Sound Source Positioning Using Vector Based Amplitude Panning,” *J. Audio Eng. Soc.*, vol. 45, pp. 456–466, 1997.

[12] H. Lee, *“The Relationship between Interchannel Time and Level Differences in Vertical Sound Localisation and Masking,” presented at the 131st Convention of the Audio Engineering Society* (2011 Oct.), convention paper 8556.

[13] R. Y. Litovsky, B. Rakerd, T. C. . T. Yin, and W. M. Hartmann, “Psychophysical and Physiological Evidence for a Precedence Effect in the Median Sagittal Plane,” *Am. Physiol. Soc. 77*, pp. 2223–2226, 1997.

[14] H. Stenzel, U. Scuda and H. Lee, “Localization and Masking Thresholds of Diagonally Positioned Sound Sources and Their Relationship to Interchannel Time and Level Differences,” In *2nd International Conference on Spatial Audio, 2014.*

[15] D. R. Begault, “Audible and inaudible early reflections: Thresholds for auralization system design,” in AES 100th Convention, Preprint 4244, 1996.

[16] E. S. Olive and F. E. Toole, “The Detection of Reflections in Typical Rooms,” *J. Audio Eng. Soc.*, vol. 37, pp. 539–553, 1989.

[17] R. Wallis and H. Lee, *“Investigation into Vertical Stereophonic Localisation in the Presence of Interchannel Crosstalk,” presented at the 136th Convention of the Audio Engineering Society* (2014 Apr.), convention paper 9026.

[18] 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.

[19] S. P. Lipshitz, “Stereo Microphone Techniques: Are the Purists Wrong?” *J. Audio Eng. Soc.,* vol. 34, no. 9, pp. 717–743 (1986 Sep.).

[20] K. Hamasaki, “Multichannel recording techniques for reproducing adequate spatial impression,” *Proc. of the Audio Engineering Society 24th International Conference* (2003).

[21] H. Lee and C. Gribben, “Effect of Vertical Microphone Layer Spacing for a 3D Microphone Array,” *J. Audio Eng. Soc.*, vol. 62, no. 11 (2014 Dec.).

[22] U. Scuda, H. Stenzel and H. Lee, “Perception of Elevated Sound Image Recorded with 3D-Audio Microphone Arrays,” In *2nd International Conference on Spatial Audio*, 2014.