

INVITED PAPER

Subjective evaluations of three-dimensional, surround and stereo loudspeaker reproductions using classical music recordings

Callum Eaton and Hyunkook Lee*

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

(Received 8 October 2021, Accepted for publication 5 December 2021)

Abstract: The present study subjectively evaluated loudspeaker reproductions of four different classical recordings in 0+2+0 (stereo), 0+5+0 (surround), 4+5+0 (surround with four height channels), each of which was downmixed from the original 9+10+3 (i.e. NHK 22.2), in terms of four attributes: listener envelopment (LEV), presence (i.e. sense of being there), overall tonal quality (OTQ) and overall listening experience (OLE). Prior to the main experiment, the playback levels of the upper and bottom loudspeaker layers relative to the middle layer level were subjectively adjusted for each of the original 9+10+3 recordings. It was found that the preferred levels of the upper and bottom layers were around 4 dB and 6 dB lower than that of the middle layer, on average. From multiple comparison listening tests, the perceived degradation from the original 9+10+3 to 4+5+0 was found to be significantly dependent on the recording technique used as well as the programme material. It was also found that 0+5+0 was not significantly different from 4+5+0 in general. Overall, LEV was most correlated with OLE, whilst Presence and OTQ tended to have a strong association.

Keywords: 3D audio, Loudspeaker format, Downmixing, Overall listening experience, Microphone technique

1. INTRODUCTION

1.1. Background

With the recent advance of production and transmission technologies and the increasing number of commercial services, three-dimensional (3D) audio is currently receiving much attention from content creators and consumers as well as researchers and developers. For example, Dolby Atmos [1] is an object-based 3D audio format that was originally introduced for cinema, but is now gaining a popularity for music production too. MPEG-H [2] is a recently standardised audio codec that can store or transmit channel-based, object-based or scene-based 3D audio content in an efficient way. It is not only used for broadcasting, but also for SONY's 360 Reality Audio format [3]. An increasing number of record labels and music streaming service platforms are now producing and delivering 3D audio content produced in the aforementioned formats.

Despite the availabilities of object-based audio and scene-based audio (e.g. Ambisonics) technologies, 3D recordings made for classical music are still largely channel-based. This is related to the nature of the music

and the recording tradition. That is, there is typically no need to automate movements of individual objects (e.g. spot microphones for certain instruments). The so-called “height” channels in a 3D loudspeaker format are mainly used to add additional ambience to provide a more realistic and enveloping sound field in reproduction. Furthermore, recording engineers use main and ambience microphone arrays for capturing the overall acoustic scene or/and multiple spot microphones. Typically, each of the main and ambience array signals is discretely routed to each corresponding loudspeaker, whereas the spot microphone signals are panned between different channels. Over the last decade, a number of different microphone arrays for 3D acoustic recording have been developed; an extensive review is provided in [4].

For the reproduction of a 3D recording over loudspeakers, Recommendation ITU-R BS.2051 [5] specifies a number of different channel and loudspeaker configurations and recommends a standardised labelling and naming conventions for them, which are used throughout this paper. The convention identifies the number of channels in the form of “upper layer + middle layer + bottom layer.” The largest configuration specified in the document is 9+10+3, which is widely known as “NHK 22.2.” Due to its large channel count, it is often regarded as a reference configuration that can deliver a greater sense of spatial

*e-mail: h.lee@hud.ac.uk
[doi:10.1250/ast.43.149]

impression and a better overall quality than the smaller configurations such as 4+5+0 (a.k.a. 9.1 or 5.1.4) or 0+5+0 (a.k.a. 5.1). There exist various microphone techniques developed for 9+10+3 recording, e.g. [6–9]. However, when a 9+10+3 content is delivered to the end user, especially in a home environment, there might arise the need to downmix it to a smaller loudspeaker configuration. In this case, it would be of interest to see how much, if any, of quality degradation would occur.

1.2. Limitation of Previous Research

Silzle *et al.* [10] compared different loudspeaker configurations in terms of basic audio quality (BAQ) in the context of downmixing (9+10+3 vs. 4+5+0 vs. 0+5+0 vs. 0+2+0), using various types of 3D audio content made for 9+10+3. They conducted two sets of multiple comparison tests; with and without 9+10+3 as a reference. When 9+10+3 was a reference, which was fixed at 100 in the scale ranging from 0 to 100, it was found that the downmixed smaller formats were lower than 9+10+3 in BAQ for all conditions. However, when 9+10+3 was included as one of the test stimuli, a number of 4+5+0 downmix conditions had no significant difference from their original 9+10+3 versions.

Schoeffler *et al.* [11] also compared 3D, surround and stereo loudspeaker reproductions, in the context of downmixing. Their focus was to examine difference in ratings between BAQ and overall listening experience (OLE). OLE is essentially the quality of experience (QoE) in the context of audio. The BAQ test used the 3D as a reference, whereas the OLE test evaluated each format either individually or in a multiple comparison without a reference. Their findings were consistent with those by Silzle *et al.* [10]’s described above; there was no significant difference between 3D and 2D overall.

Although the previous studies provide a useful insight about the perceived degradation of downmixing from 9+10+3 to a smaller format, they are limited in two aspects. Firstly, they used a variety of test items of different genres, e.g. pop, rock, sound effects, outdoor soundscape, classical and movie. However, there were only two classical music items of a similar type included in both studies. Therefore, it is difficult to know if the results could be applied to different types of classical music recording made in different acoustic spaces. Furthermore, the recording techniques used for those test items were not described in their papers, but perceived degradation due to downmixing may be associated with the characteristics of signals captured using different microphone configurations.

Secondly, Silzle *et al.* [10] and Schoeffler *et al.* [12] focused on the assessment of “overall” quality or experience, rather than specific perceptual attributes. Although the BAQ and OLE evaluations provide useful insights into

how different loudspeaker formats compare, the potential reasons for the rating results are not known. Those global attributes might be associated with specific lower-level attribute such as listener envelopment (LEV) and tonal quality. Furthermore, depending on the attribute, the magnitudes of perceived differences among the different formats may vary.

1.3. Aims of the Present Study

From the above background, the present study aims to evaluate the perceived degradation of 9+10+3 classical recordings downmixed to 4+5+0, 0+5+0 and 0+2+0 formats, in terms of four different attributes: Listener Envelopment (LEV), Overall Tonal Quality (OTQ), Presence (i.e. sense of “being there”) and Overall Listening Experience (OLE). Four recordings made using two broadly different microphone techniques in acoustic spaces were used as test items. They comprised high-quality acoustic recordings of a brass band, a string quartet, a large orchestra and a choir with an orchestra. It was of main interest of the study to examine how the perceived differences among the formats would depend on the test item and perceptual attribute. Furthermore, the correlations among those attributes in ratings were examined.

Prior to the main experiment, an initial experiment was conducted to elicit subjectively preferred playback levels of the upper and bottom layers to a fixed middle layer level for the recordings in 9+10+3 reproduction. The resulting levels were to be used in the later main experiment. This experiment was considered to be the necessary first step since the initial upper layer level may affect the downmix quality. There has been a lack of experimental data on the optimal playback level balance among the middle, upper and bottom layers in 3D classical recording. King *et al.* [12] investigated the subjectively preferred level of the upper layer for orchestra, piano and string trio recordings, and found that the result varied between -9.6 dB and -19.3 dB below the level set by the original mix engineer, depending on the recording. This indicates there tends to be a considerable discrepancy between the content producer’s and the listener’s preferences on the upper layer level. However, their study did not report the level difference between the middle and upper layers in the original mix; thus, it is not possible to know the preferred upper layer level relative to the middle layer level. In contrast, the present study measured the preferred level differences of the upper and bottom layers to the middle layer in terms of sound pressure level (SPL) at the listening position; thus, the results might be transferrable for a general layer balancing purpose.

The rest of the paper is organised as follows. Section 2 describes Experiment 1: optimal playback levels of the upper and bottom layers. Section 3 details Experiment 2:

evaluation of loudspeaker formats downmixed from 9+10+3. Finally, Sect. 4 provides general discussions on the main findings from the studies and proposes further research.

2. EXPERIMENT 1

This experiment determined the subjectively preferred playback levels of the upper and bottom layers relative to the middle layer, primarily to achieve the most optimal mix balances among the three layers for Experiment 2. Additionally, comments from participating subjects were collected to gain insight into the rationales for their level judgements for each of the upper and bottom layers.

2.1. Physical Setup

Both Experiment 1 and Experiment 2 were conducted in a Recommendation ITU-R BS.1116-3 [13]-compliant listening room at the Applied Psychoacoustics Lab of the University of Huddersfield (6.2 m × 5.2 m × 3.5 m, RT = 0.25 s, NR = 12). A total of 22 loudspeakers were used in this experiment (7 Genelec 8331As and 15 Genelec 8040As). They were placed in the arrangement based on the Recommendation ITU-R BS.2051-2 [5], as summarised in Table 1, and visually hidden using acoustically transparent curtains. Subwoofer channels were not utilised in this study, as is often the case for classical music recording.

Table 1 9+10+3 loudspeaker configuration and labels, based on Recommendation ITU-R BS.2051-2 [5].

Spk. Label	Ch. Label	Ch. Name	Loud- speaker positions (Azimuth°, Elevation°)
M+060	FL	Front left	60, 0
M-060	FR	Front right	-60, 0
M+000	FC	Front centre	0, 0
M+135	BL	Back left	135, 0
M-135	BR	Back right	-135, 0
M+030	FLc	Front left centre	30, 0
M-030	FRc	Front right centre	-30, 0
M+180	BC	Back centre	180, 0
M+090	SiL	Side left	90, 0
M-090	SiR	Side right	-90, 0
U+045	TpFL	Top front left	45, 45
U-045	TpFR	Top front right	-45, 45
U+000	TpFC	Top front centre	0, 45
T+000	TpC	Top centre	0, 90
U+135	TpBL	Top back left	135, 45
U-135	TpBR	Top back right	-135, 45
U+090	TpSiL	Top side left	90, 45
U-090	TpSiR	Top side right	-90, 45
U+180	TpBC	Top back centre	180, 45
B+000	BtFC	Bottom front centre	0, -30
B+045	BtFL	Bottom front left	45, -30
B-045	BtFR	Bottom front right	-34, -30

The audio signals were reproduced via a Merging Horus D/A converter.

2.2. Test Items

30-second-long excerpts from four sets of acoustic musical recordings made for the 9+10+3 loudspeaker reproduction were used as test items for the current study. They are labelled as Brass Band, Choir, Orchestra and String Quartet. These recordings were chosen not only because they represent different types of musical ensemble, but also because they were made at a professional quality using different microphone configurations and mixing approaches. Brass Band and String Quartet were recorded by the second author of this paper. Choir and Orchestra were recorded by Toru Kamekawa and Will Howie, respectively, and the original multitrack sources were acquired directly from the recording engineers. The technical details of the recordings are summarised in Appendix.

2.3. Subjects

Seven experienced subjects participated in this test. They comprised two staff researchers and five post-graduate researchers at the Applied Psychoacoustics Lab (APL) of the University of Huddersfield. All of them had experiences and in-depth knowledge in music production and spatial audio evaluation. The two staff researchers had at least five years of experiences in 3D recording and mixing with height channels. Two of the post-graduate researchers had at least one year of experience in the same. They all reported to have normal hearing.

2.4. Test Method

Only The subjects were asked to mix the upper and bottom layers to the levels that were deemed most preferable in relation to the middle layer level, which was fixed at the SPL of 74 dB in L_{Aeq} at the listening position. The Reaper digital audio workstation (DAW) was used for the playback of the recordings. Separate group tracks for the upper and bottom layer signals were created, and their levels were adjusted using faders of a Korg nanoKontrol 2 controller that was connected to the DAW. The DAW was not visible to the subjects so that they could focus on the level adjustment task without a potential visual bias.

For each excerpt, the faders for the upper and bottom layers were initially placed at the bottom position (i.e. -inf), thus starting with only the middle layer playing audio. This was to prevent the subjects from being biased by a starting fader position and the resulting loudness. After each subject completed the task, the playback SPL for each of the upper and bottom layers was measured at the listening position. As mentioned earlier, King *et al.* [12] recorded

the preferred level of the upper layer in 4+5+0 reproduction and reported its difference to the mix level set by the original engineer for the recording. However, in the present test, the aim was to obtain the differences of the upper and bottom layers to the middle layer in terms of playback SPL, which would allow for a better understanding about subjectively preferred level balances among the layers.

The Subjects were allowed to mute and solo the three layers in the mixing process, and to listen to the recordings as long as they wanted. Head movement was not restricted but not encouraged. The subjects were also instructed to remain seated where they were initially positioned in the sweet spot. Each programme material was presented in a randomised order to each subject.

In addition, the subjects were asked to freely describe the reasons for their mix level decision on a word document after completing the task in each trial. This was to gain insights into perceptual features that may be linked to the preferred mix levels of the layers.

2.5. Results and Discussion

2.5.1. Level balance between the layers

The SPLs of the bottom and upper layers resulting from the preferred mix level determined by each subject were collected and analysed statistically using R. Shapiro-Wilk test was first performed to test the normality of the data. Although it was found that all conditions except the bottom layer for Orchestra ($p = 0.008$) had a normal distribution of the data, it was decided to use Wilcoxon test, which is a non-parametric statistical method, for a significance testing, due to the relatively low number of data points.

Figure 1 plots the median and associated notch edge (non-parametric 95% confidence interval) for each source-

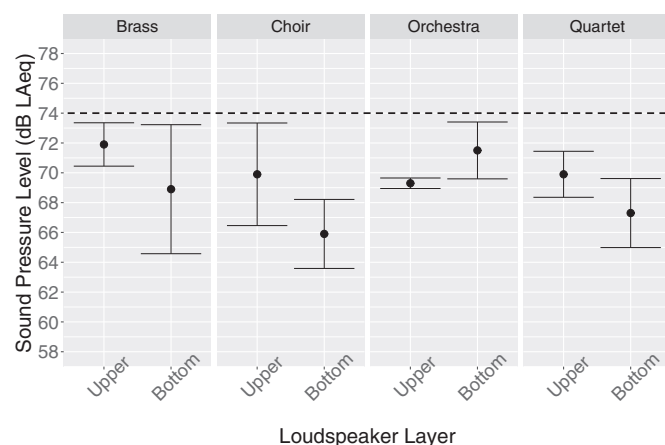


Fig. 1 Medians and associated non-parametric 95% confidence intervals for the sound pressure levels (SPLs) of the preferred bottom- and upper-layer playback levels relative to the middle layer level. The dotted line indicates the SPL of the middle layer, which was fixed at 74 dB in L_{Aeq} .

Table 2 Median preferred playback levels (SPLs) of the upper and bottom layers for each test item.

Test item	Upper layer (dB in L_{Aeq})	Bottom layer (dB in L_{Aeq})
Brass band	71.9	66.5
Choir	69.9	65.9
Orchestra	69.3	71.1
String quartet	69.9	67.3
Average	70.3	67.7

layer condition. The horizontal dotted line crossing all four conditions indicate the level of the middle layer level that was fixed at 74 dB in L_{Aeq} . The plots therefore show how the subjects mixed the upper- and bottom-layer levels in relation to middle layer level.

It is apparent from the plots that, for all four test items, the median values for both the upper and bottom layers did not exceed the middle layer level. Wilcoxon test confirmed that the SPLs of all conditions were significantly lower than that of the middle layer ($p < 0.05$), apart from the upper layer of Choir.

The median SPLs are summarised in Table 2. These values were used to calibrate the playback levels of the bottom and upper layers in Experiment 2. For the bottom layer, the median L_{Aeq} values ranged from 65.9 dB to 71.1 dB, with Orchestra having a considerably higher level than Choir and Quartet with a large effect size ($r = 0.52$ and 0.46 , respectively) despite Wilcoxon test with Bonferroni correction suggesting non-significant differences among those conditions. On the other hand, the upper layer results varied substantially less (69.3 dB to 71.9 dB); no pairs of conditions had a significant difference or a large effect size.

Comparing between the bottom and upper layers for each program material, Brass, Choir and Quartet had the median SPL of the upper layer being 2.6 dB to 5.4 dB higher than that of the bottom layer. On the other hand, for Orchestra, the upper layer was 1.8 dB lower than the lower layer in SPL.

A potential reason why Orchestra had a substantially greater SPL than Choir and String Quartet for the bottom layer is explained as follows. The bottom layer signals of Brass Band and String Quartet sounded highly reverberant since the cardioid microphones used for the layer faced away from the ensembles to reject direct sounds. On the other hand, the Orchestra recording used cardioid microphones that were tilted about 45° downwards, pointing towards the floor and the first row of the orchestra. Therefore, the bottom layer had strong early reflections from the floor, combined with some direct sounds. It is deemed that raising the levels of the reverberant bottom layer signals for Brass Band and String Quartet too high might have risked the perception of “muddiness” and “lack

of clarity,” which were reported by some subjects in their additional comments. However, the floor reflections might have been found to be more useful to create a more realistic sound field in the 9+10+3 reproduction.

For Choir, the bottom layer microphones were originally spot microphones used for the woodwind section. Therefore, the direct sound levels of the instruments were high in the signals. Considering the frequency range of the woodwind instruments focused around middle frequencies, the bottom layer signals might have had an interference with those of the same instruments captured by the main microphone array (Decca Tree) for the middle layer in terms of tonal balance and localisation.

In terms of the differences among the test items in the upper layer SPLs, Brass Band had around 3 dB higher SPL than that of Orchestra, which was significant ($p < 0.05$). The upper layer microphones used for Brass Band were vertically oriented so that direct sounds (i.e. inter-channel crosstalk) would be suppressed maximally and diffuse reverberant sounds would be captured mainly. On the other hand, some of the microphones used for the upper layer in the Orchestra recording would have captured substantially more direct sounds due to their polar patterns and orientations. This might explain why the upper layer for Brass Band was higher in SPL than that for Orchestra.

2.5.2. Comments from the subjects

In addition to the mixing task, comments were collected from the subjects to better understand the reasons behind their level balance decisions and to provide insight into the salient perceptual attributes in mixing the different layers.

Based on the semantics of the comments obtained, it was possible to group them in eight categories, referring to:

- Listener envelopment.
- Vertical image position/spread.
- Horizontal image position/spread.
- Reverberance.
- Tonal quality/naturalness.
- Overall experience.
- Sense of presence/being there.
- Miscellaneous comments, e.g. sound energy, level balance, and environmental depth.

Figure 2 plots the percentage of comments in each category for the upper and bottom layers, respectively. As can be seen in Fig. 2, for the upper layer, the largest proportion of comments were focused around the topic of listener envelopment (36%), with tonal quality (21%) and vertical spread or source elevation (17%) also being prominent in the reasoning for choosing the mix level of the layer. For the bottom layer, on the other hand, tonal quality (40%) was the most prominently commented. Interestingly, no references were made regarding overall experience of the mix or reverberance.

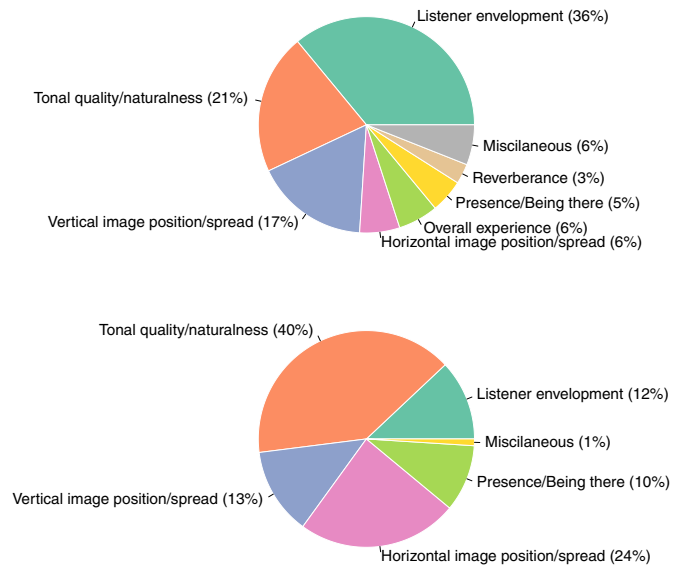


Fig. 2 Proportions of the categories of comments given by the subjects with regard to their rationales for the mixing balance decisions for the upper layer (the upper pie chart) and bottom layer (the lower pie chart).

That the upper layer mix had a higher percentage of comments regarding listener envelopment than the bottom layer mix seems to be down to the fact that the 9+10+3 loudspeaker configuration has nine upper layer loudspeakers surrounding the listener in a circular arrangement against only three bottom layer loudspeakers placed in front of the listener.

It is interesting to observe that the bottom layer mix produced more comments on the tonal quality/naturalness. The subjects remarked that with an increased level of the bottom layer, instruments often sounded more natural though with too much level they could sound ‘muddy,’ thus requiring a careful level balance with the middle layer.

The bottom layer also seems to have had a larger effect on horizontal image position and spread than the upper layer. This might be explained as follows. The left and right loudspeakers in the bottom layer were placed at $\pm 45^\circ$, whereas those in the middle layer were at $\pm 30^\circ$. Furthermore, research suggests that due to the so-called “pitch-height” effect [14,15], middle frequencies around 500 Hz to 2 kHz would be localised around the listener’s ear height, whereas lower frequencies tend to be perceived below the ear height, regardless of the physical height of the loudspeaker. Based on this, it could be suggested that the middle frequency components of the bottom layer signals contributed to the widening of the perceived position and spread of the source images without introducing a downward image shift, whilst the low frequency components caused a vertical image spread towards the floor level.

3. EXPERIMENT 2

This experiment investigated the influence of spatial downmixing on four perceptual attributes: Listener Envelopment (LEV), Presence, Overall Tonal Quality (OTQ), Overall Listening Experience (OLE). The original 9+10+3 recordings were downmixed to the 4+5+0, 0+5+0 and 0+2+0 formats. The experiment was conducted in the same Recommendation ITU-R BS.1116-compliant listening room used in Experiment 1. This section describes the downmixing method, listening test procedure, followed by the presentation and discussion of the results.

3.1. Test Items and Loudspeaker Formats

The same 9+10+3 format recordings from Experiment 1, described in Sect. 2.1, were used in this experiment. The levels of the upper and bottom layers were calibrated to the median L_{Aeq} values found in Experiment 1 (see Table 3). The original recordings were downmixed to the 4+5+0, 0+5+0, 0+2+0 and 0+1+0 formats, using a downmixing algorithms proposed by Silzle *et al.* [10]. This was chosen over an active (i.e., adaptive) algorithm (e.g. [16]) for two reasons. Firstly, it was aimed to apply the same downmixing strategy consistently across all of the recordings and examine how the ratings of the four different formats vary depending on the recording technique used. As mentioned above, Brass Band and String Quartet were recorded using a microphone technique that was optimised in terms of suppressing interchannel crosstalk, whereas Orchestra and Choir used techniques that would likely allow stronger interchannel crosstalk in the upper and bottom layers. Secondly, Silzle *et al.* [10]'s

study also compared the same formats in terms of overall sound quality, and using the same algorithms to their study would allow for a comparison with the results for other attributes obtained from the current experiment. The mono 0+1+0 format was included as a low anchor, which was expected to be rated the lowest. However, it was of interest to see how the stereo (0+2+0) format would be rated over the surround and 3D formats depending on the attribute. Table 3 summarises the downmix algorithms used for each downmixed format.

Following the downmixing process, the overall playback levels of the original recording and each downmixed version were calibrated to 77.5 dB in L_{Aeq} .

3.2. Subjects

Eleven subjects took part in this experiment. They comprised the two staff and five post-graduate researchers who participated in Experiment 1 and four final year undergraduate students from the Music Technology courses at the University of Huddersfield. As mentioned already, the staff and post-graduate researchers had much experience in spatial audio production and evaluation. The final year students had at least one year of critical listening training and spatial audio recording, and had prior experiences with various types of spatial audio evaluation in controlled listening test settings. They all reported to have normal hearing. Due to the nature of the listening test requiring critical listening skills, it was deemed to be sufficient to use those eleven selected subjects rather than adding more unexperienced subjects. To increase the number of observations, however, each subject tested each experimental condition twice.

Table 3 Downmix algorithms for 9+10+3 to different formats and loudspeaker positions.

Format	SP label	Channel label	Loudspeaker positions (Azi°, Ele°)	Downmix algorithms
4+5+0	M+030	L	30, 0	FL + 0.7071(FLc+SiL) + BtFL
	M-030	R	-30, 0	FR + 0.7071(FRc+SiR) + BtFR
	M+000	C	0, 0	FC + 0.7071(FL+FR) + BtFC
	M+110	Ls	110, 0	BL + 0.7071(SiL+BC)
	M-110	Rs	-110, 0	BR + 0.7071(SiR+BC)
	U+030	Ltf	45, 45	TpFL + 0.7071(TpSiL+TpFC) + 0.5(TpC)
	U-030	Rtf	-45, 45	TpFR + 0.7071(TpSiR+TpFC) + 0.5(TpC)
	U+110	Ltr	135, 45	TpBL + 0.7071(TpBC+TpSiL) + 0.5(TpC)
	U-110	Rtr	-135, 45	TpBR + 0.7071(TpBC+TpSiR) + 0.5(TpC)
0+5+0	M+030	L	30, 0	FL + 0.7071(FLc+SiL) + BtFL + TpFL + 0.7071(TpSiL)
	M-030	R	-30, 0	FR + 0.7071(FRc+SiR) + BtFR + TpFR + 0.7071(TpSiR)
	M+000	C	0, 0	FC + 0.7071(FL+FR) + BtFC + TpFC + 0.7071(TpC)
	M+110	Ls	110, 0	BL + 0.7071(SiL+BC)
	M-110	Rs	-110, 0	BR + 0.7071(SiR+BC)
0+2+0	M+030	L	30, 0	L + 0.7071(C) + 0.7071(Ls)
	M-030	R	-30, 0	R + 0.7071(C) + 0.7071(Rs)
0+1+0	M+000	C	0, 0	0.7071(L + R)

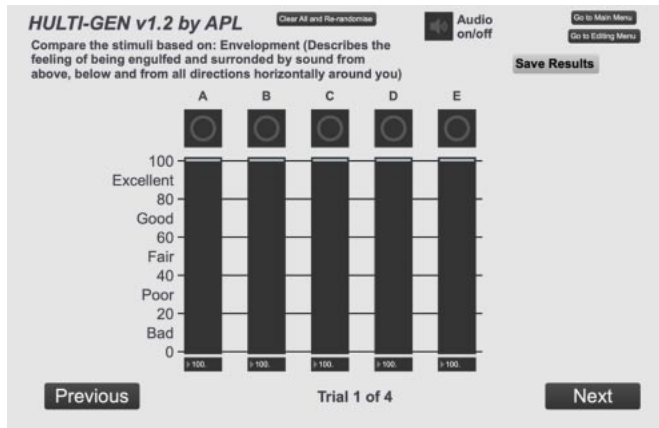


Fig. 3 Graphical user interface used for Experiment 2.

3.3. Test Method

For the listening test, the HULTI-GEN software [17] was used to create a graphical user interface that provided the subjects with five sliders and corresponding play buttons (see Fig. 3). Each of the sliders represented one of the five format stimuli for each of the four recordings, which were time-synchronised and looped. The order of the stimuli was randomised for each subject and each trial. Each slider had a range of 0 to 100 with semantic labels based on those outlined in the Recommendation ITU-R BS.1534-3 specification [18].

The subjects were sat in the central position (i.e. the sweet spot), with their ear height adjusted to be at the height of the acoustic centre of the middle layer loudspeakers. They were not restricted in terms of head movement, but not encouraged to do so. Their task was to compare the five different formats and rate them using the sliders. They could listen to the stimuli as many times as they wanted. Each trial tested one of the four stimuli at a time, either the Brass Band, Orchestra, String Quartet or Choir recordings. The trials were presented in a randomised order, with each trial repeated twice.

It is worth noting that the original 9+10+3 stimuli were not presented as a reference. This would have forced the subjects to assume the stimuli to be rated at 100 as in a MUSHRA (Multiple Stimulus with Hidden Reference and Anchor) test [18]. However, previous research by Silzle *et al.* [10] and Schoeffler *et al.* [11] showed that there is a significant difference between tests with and without a 9+10+3 reference in terms of the ratings given to different loudspeaker formats. For example, 9+10+3 recordings were not necessarily higher than 4+5+0 in ratings, depending on the programme material. Both studies proposed that a reference should not be used for evaluating overall sound quality [10] or overall listening experience (OLE) [11].

Four attributes were chosen for the test and defined as follows.

- Listener Envelopment (LEV): The feeling of being engulfed and surrounded by reverberant sound from above, below and all directions horizontally around the listener (based on [19,20]).
- Presence: The sense of ‘being there’ (i.e. in the recording venue).
- Overall Tonal Quality (OTQ): Any features of sound related to timbral aspects, e.g. spectral balance, clarity, phasiness, etc. Not to be confused with any spatial attributes.
- Overall Listening Experience (OLE): The level of overall satisfaction towards an audio content, i.e. quality of experience (QoE) in the context of listening (based on [11,21]).

LEV was chosen as it is one of the most widely referred spatial attributes in surround sound reproduction as well as in concert hall acoustics. It was of interest to see how the upper and bottom loudspeaker layers contribute to LEV in 3D reproduction. It is important to note that although LEV is often associated with immersive auditory experience, they should not be used as synonyms as pointed out in [22]. LEV might be one of the determinants for a higher-level experience-based attributes such OLE and immersive experience.

The Presence attribute is one of the main pillars of immersive experience [23], and can be categorised into physical presence, social presence and self-presence in virtual reality [24]. In the context of the present study, the focus was given only on the physical presence, which is essentially about a sense of being in the recording venue.

OTQ was also included since it was considered that there might be a timbral change when the upper and bottom layer signals are downmixed to a smaller format depending on the levels of direct sounds captured by the corresponding microphones. In the context of surround and 3D acoustic recordings, the direct sounds captured by surround and height channels could be considered as interchannel crosstalk, which might cause a spectral distortion in the ear-input signals [25] as well as an interference in horizontal and vertical image localisation, respectively [26,27].

3.4. Results and Discussions

The data collected from the listening test were plotted and statistically analysed using R. Shapiro-Wilk test found that not all experimental conditions had a normal data distribution. Therefore, medians and associated notch edges (non-parametric 95% confidence intervals) are plotted for each recording and each attribute in Fig. 4. As can be seen from the Fig. 4, the overall spread of ratings among the different formats was substantially large. Friedman test confirmed that the main effect of loudspeaker format was significant for all test items and attributes ($p < 0.001$). The low anchor 0+1+0 stimuli were mostly

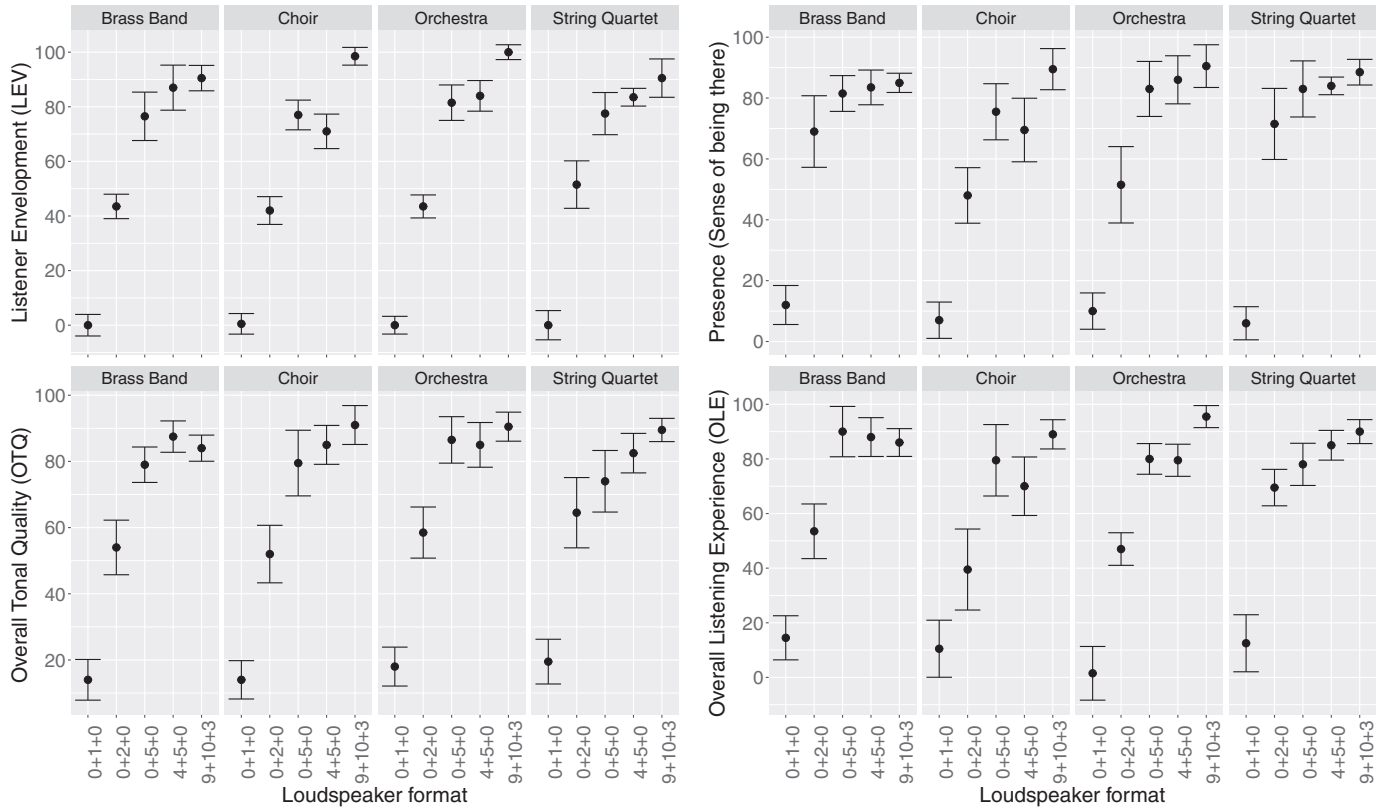


Fig. 4 Medians and associated notch edges (non-parametric 95% confidence interval) of the listening test data, plotted for different recordings and attributes.

rated in the “poor” category (0–20) as expected. From 0+2+0 to 9+10+3, the categories of the median ratings ranged from “fair” (40–60) to “excellent” (80–100). However, the increment from a smaller format to a larger format did not always produce a monotonic increase in median ratings. The significances of differences among the formats are analysed using Wilcoxon signed rank test with Bonferroni correction. The following sections present and discuss the differences among the different formats, their dependency on the test item and attribute, and the correlations among the attributes.

3.4.1. 4+5+0 vs. 9+10+3

Firstly, the difference between the original 9+10+3 and its downmixed 4+5+0 was dependent on the recording and attribute. That is, 4+5+0 was found to be significantly lower than the 9+10+3 only for Choir and Orchestra in terms of OLE ($p = 0.012$ and 0.008 , respectively) and LEV ($p = 0.0005$ and 0.0013 , respectively). Brass Band and String Quartet produced no significant differences between the two formats for all attributes tested ($p > 0.05$). Furthermore, comparing amongst all test items, Choir and Orchestra have generally lower median ratings than Brass Band and String Quartet for 4+5+0 in OLE and LEV, even though the opposite is true for 9+10+3.

These results may be not only due to the differences in the ensemble size and musical scale, but also related to the

microphone techniques used for the recordings. Brass Band and String Quartet were recorded in the same concert hall using PCMA-3D-based microphone arrangements, where the middle- and upper-layer microphones were arranged in the horizontally spaced and vertically coincident (HSVC) fashion [4,9]. On the other hand, Choir and Orchestra were recorded using horizontally and vertically spaced (HVS) approaches [4,7], with substantially larger distances between microphones than Brass Band and String Quartet. The HSVC microphone arrangement avoids vertically oriented interchannel time differences whilst sufficiently reducing the level of interchannel crosstalk (i.e. direct sounds) captured in the upper layer signals [4,28]. Therefore, it aims to maximise the separation between direct sounds in the frontal channels of the middle layer and ambient sounds in the side, rear and top channels. On the other hand, the microphone arrangements used for Choir and Orchestra seem to allow more interchannel crosstalk in the upper layer signals due to the polar patterns of the microphones and their orientations.

Based on this, it is hypothesised that Choir and Orchestra might have substantially lost the horizontal spread and spatial resolution of source-related images when the original nine and ten channel signals were downmixed to four and five channels respectively for the upper and middle layers. This might have resulted in the reduction of

LEV, and this seems to have also had a major influence on OLE; as will be shown in Sect. 3.4.4. OLE was found to be most correlated with LEV. Although LEV is typically defined to be reverberation-related rather than source-related [19], it could be difficult to separate reverberation from source in ongoing parts of an orchestral piece. Therefore, interchannel crosstalk fused with reverberation could also contribute to LEV. On the other hand, Brass Band and String Quartet might have had less noticeable reduction of LEV since the downmixed signals are mostly of diffuse reverberant sounds, for which would likely be more difficult to perceive a spatial gap between loudspeakers than direct sounds.

In addition, these results also seem to contrast the findings from Silzle *et al.* [10]’s experiment, which used the same downmix algorithm used in this study. Their results suggested that 9+10+3 (original) and 4+5+0 (downmixed) did not produce a significant difference in Overall Sound Quality (OSQ) for all orchestral recordings tested when there was no hidden reference among the test stimuli. However, the present results here suggest that a degradation from 9+10+3 to 4+5+0 could be significant for orchestral recordings in OLE and LEV, potentially depending on the microphone techniques used as discussed above.

3.4.2. Surround vs. 3D formats

Schoeffler *et al.* [11] found in their format comparison experiments that the increment from surround (0+5+0) to 3D audio (0+9+0 and 9+10+3) was smaller in OLE ratings compared to than in basic audio quality (BAQ) ratings. Silzle *et al.* [10] also found that 4+5+0 was not significantly higher than 0+5+0 for classical music recordings. The results obtained in the present study shows a similar trend; all experimental conditions, apart from Brass Band for LEV ($p = 0.005$) and String Quartet for OLE ($p = 0.047$), were found to have no significantly higher median of 4+5+0 than 0+5+0 ($p > 0.05$). The fact that the four upper layer loudspeakers added to 0+5+0 did not produce any improvement in LEV is in line with observations made in [25]; the addition of the upper layer to the middle layer in a 4+5+0 reproduction did not cause a meaningful change in the interaural cross-correlation coefficient (IACC) of binaural room impulse responses resulting from the reproduction of 3D microphone array signals.

However, it is worth noting that the test items used in [25] as well as the present study were classical recordings with the upper layer signals consisting of ambient sounds mostly. Such ambience reproduced from the upper layer might not provide a strong sense of vertical image location and spread due to its frequency spectrum having a roll-off at high frequencies. Based on the pitch-height effect [14,15], low to middle frequency components would tend to be localised at the listener’s ear height or below.

Therefore, it can be considered that the impact of adding the ambient upper layer would be low in terms of vertical LEV, but might be rather on perceived image distance due to a decrease in direct-to-reverberant energy ratio (DRR). The upper layer, however, may have different impacts on LEV and other attributes for different content types, such as film and popular music where audio “objects” could be vertically panned.

Furthermore, although the studies by Silzle *et al.* [10] and Schoeffler *et al.* [11] showed that 9+10+3 was always rated significantly higher than 0+5+0 in terms of OLE and OSQ, respectively, the present results indicate that 0+5+0 could be as good as 9+10+3 depending on the recording and attribute under test. That is, all test items did not have a significant difference between the two formats for OTQ and Presence ($p > 0.05$). This indicates the upper and bottom layers were not the determinant of the overall tonal quality and the sense of “being there,” and that the downmix of 9+10+3 into a conventional surround format would suffice for the specific purposes. For OLE and LEV, on the other hand, 0+5+0 was found to be significantly lower than 9+10+3 for all test items apart from Brass Band (for OLE) and String Quartet (for LEV). This seems to suggest that OLE and LEV benefit from the addition of both the upper and bottom layers as well as more surrounding loudspeakers in the horizontal plane (i.e. $9+10+3 > 4+5+0 > 0+5+0$).

3.4.3. Stereo vs. surround and 3D formats

One might expect that the stereo format (0+2+0) would naturally be rated significantly lower than the surround and 3D formats due to the lack of surrounding channels. This was indeed the case for most of the attributes and test items, but not all. That is, for Presence, the 0+2+0 versions of Brass Band and String Quartet were found to have no significant difference to 0+5+0 ($p > 0.05$). For Brass Band in particular, 0+2+0 had no significant differences even to 4+5+0 and 9+10+3 in Presence ($p > 0.05$).

This is interesting as the two test items were recorded in the same venue (St.Paul’s concert hall, $RT = 2.1$ s) using a similar microphone arrangement based on PCMA-3D [4,9]. A potential explanation for this is provided as follows. As mentioned above, the PCMA-3D approach aims to suppress interchannel crosstalk in the surround and height channels, whilst capturing decorrelated ambient sounds. When these ambient signals are downmixed to 0+2+0, they might add more constructively without influencing the imaging of the sound sources, compared to other techniques that introduce a higher amount of interchannel crosstalk. This may eventually help maintaining the realistic sense of the space of the sound even in the stereo reproduction.

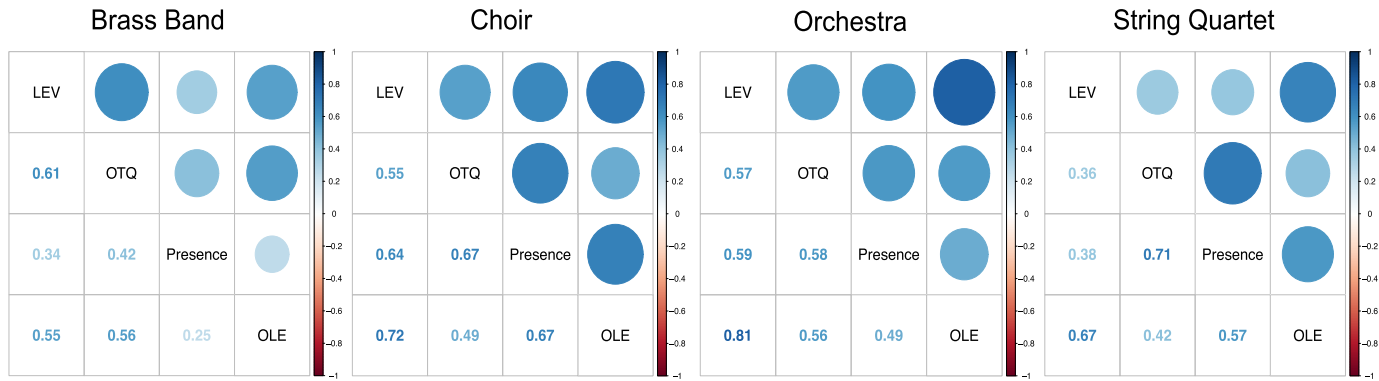


Fig. 5 Correlations among the attributes. The numbers indicate Spearman correlation coefficient (rhos).

3.4.4. Correlations among the attributes

In order to gain an insight into the relationships among the attributes tested in this study, Spearman correlation tests were performed on the data for each attribute. The 0+1+0 conditions were excluded as they were low anchors and therefore commonly rated at the bottom range of the scale. The results are shown in Fig. 5. The numbers indicate the correlation coefficients (rhos), which are visually represented as the size and colour gradient of the circle. As can be seen, the correlations among the attributes did not have a consistent pattern across the test items. For Orchestra and Choir, LEV and OLE had the highest correlation amongst all pairs of attributes ($\rho = 0.81$ and 0.72 , respectively). The other attributes all had a medium correlation. Brass Band generally had a medium to low correlation for all pairs of attributes. For String Quartet, OTQ and Presence had the highest correlation amongst all ($\rho = 0.71$).

In general, LEV was the attribute that was most correlated with OLE, with the highest ρ for Orchestra. Furthermore, for all test items apart from Brass Band, LEV had a higher correlation with OLE than OTQ. This is interesting since previous research [29] found that tonal fidelity was more important than spatial fidelity for BAQ, in the context of quality degradation by low-pass filtering of original signals as well as downmixing from 0+5+0 to 0+2+0 and 0+1+0. OTQ in the present study, however, would have been determined by potential tonal degradation due to the downmixing process alone. As shown in Fig. 4, the magnitude of degradation from 9+10+3 to 4+5+0 in OTQ was considerably smaller than that in LEV. It may be that, in the context of 3D to 2D downmixing, a change in LEV is more noticeable and related to OLE than a change in OTQ.

It is worth noting that Presence was not highly correlated with either LEV for any of the test items. Brass Band and String Quartet had a low correlation between Presence and LEV, whereas Orchestra and Choir had a medium correlation between the two attributes. It was

found that Presence was more correlated with OTQ, especially for Choir and String Quartet. It is often taken for granted that enveloping reverberation is an essential requirement for producing the sense of “being there” in surround reproduction. However, the current result seems to suggest that LEV may not be a highly important factor for Presence, potentially depending on the type of programme material, microphone technique and acoustics of the recording venue. This is also implied in the rating results for Brass Band and String Quartet, which were discussed above. That is, 0+2+0 and 0+5+0 did not have a significant difference in Presence rating, whereas Orchestra and Choir appeared to benefit from the surround channels more.

In order to gain an insight into the overall relationship among all of the attributes, principal component analysis (PCA) was performed using all data from the experiment. The biplot of PCA shown in Fig. 6 visually confirms a clear separation between the OLE/LEV group and Presence/OTQ group. The close relationship between Presence and OTQ is interesting since tonal aspect is usually not considered when discussing a sense of being there. Although this result might be specific to the context of the present study, which is downmixing, it seems to indicate that not only room-acoustical factors and spatial realism but also timbral naturalness and realism might be an important aspect to consider for creating a strong sense of being there. For instance, a heavily comb-filtered or poorly synthesised musical sound would likely be perceived to be unrealistic and implausible, thus producing a weaker sense of presence.

4. GENERAL DISCUSSION AND CONCLUSIONS

This section summarises the main findings from the two experiments conducted in this study and discuss practical implications, limitations and further research required.

Experiment 1 showed that the preferred level balances of the upper and bottom layers against the middle layer

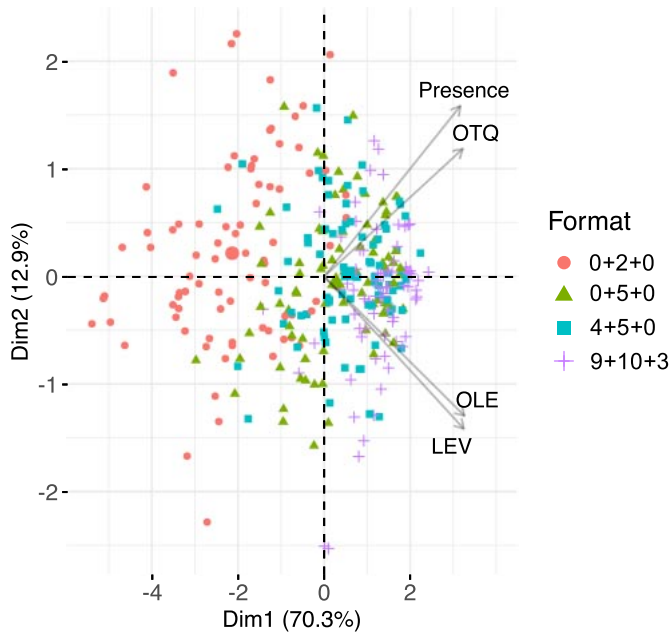


Fig. 6 Biplot of the principal component analysis performed using all data from Experiment 2.

varied depending on the test item. In particular, the Brass Band recording allowed for a lower level for the bottom layer and a higher level for the upper layer than the Orchestra recording. Brass Band had more diffuse reverberation in the bottom and upper layers, whereas Orchestra had stronger floor reflections and direct sounds in the bottom layer and stronger direct sounds in the upper layer. From this, it is hypothesised that for the bottom layer, early reflections from the floor could be more useful for the bottom layer than diffuse reverberation, whilst the upper layer might benefit more from diffuse reverberation rather than direct sounds, in the context of acoustic 3D recording made using microphones. A future experiment will aim to verify this hypothesis by the amount of direct sound, early reflections and reverberation in a more systematic way for given test items.

Experiment 2 found that the perceived degradation from the original 9+10+3 to 4+5+0 in the downmixing scenario depended on the test item and attribute. Although the test items used in this study used different microphone configurations, Brass Band and String Quartet were made using PCMA-3D-based configurations that aimed to minimise interchannel crosstalk (i.e. direct sound captured by microphones in the bottom and upper layer), whereas Orchestra and Choir recordings had a higher level of interchannel crosstalk with time delays, which would cause comb-filtering when the signals of the two layers are summed. The relatively crosstalk-free Brass Band and String Quartet recordings did not have a significant difference between 9+10+3 and 4+5+0 for all attributes. From this, it is hypothesised that the less the amount of

interchannel crosstalk in the bottom and upper layers was, the less the degree of quality degradation would be in the passive downmixing of a 9+10+3 recording. However, as the programme material was different for each test item in the present study, this hypothesis needs to be validated in a future study that will look into the influence of the amount of interchannel crosstalk on downmix quality for individual programme materials.

Experiment 2 also showed that 0+5+0 was not significantly different from 4+5+0 in general. This suggests that, in the context of classical recording made in a reverberant space, the four added “height” channel signals might not provide a major perceptual benefit. It is important to note that the overall playback levels of all test conditions were normalised. If the upper layer of 4+5+0 were turned on and off, the difference against 0+5+0 might be more audible. However, in the level-matched scenario, the current results suggest that the benefit of the upper layer might not be significant. However, since the 4+5+0 conditions were produced from the original 9+10+3 recordings via the downmixing algorithm, it is difficult to generalise this discussion. If the downmix was made directly from an original 4+5+0 recording to 0+5+0, the effect of the four upper layer signals on the perceived attributes would be examined more clearly.

The present study used the passive downmix algorithms that were used in Silzle *et al.* [10]’s study for the reasons described in Sect. 3.1. However, passive downmixing is limited in that it would depend on the characteristics of the signals. Especially, if the upper or bottom layer signals contained a high level of delayed interchannel crosstalk, there might occur audible spectral distortion when they are summed with the middle layer signals. Active downmixing can adapt to the temporal and spectral characteristics of the channel signals. For example, MPEG-H 3D audio codec utilises an active downmix algorithm that allows for phase alignment [16]. Several active downmix algorithms were proposed for surround sound [30,31]. However, more research is required on the development and evaluation of active downmixing algorithms for 3D audio.

Overall, ratings on listener envelopment (LEV) were found to be most correlated with those on overall listening experience (OLE), but not with Presence (i.e. sense of being there). This seems to suggest that although LEV is an important attribute for the quality of experience, it may not be a determinant of physical realism and plausibility in sound reproduction. It is worth noting that LEV is originally measured by the amount of lateral sound energy [32] or the correlation between the ear-input signals [19], rather than how authentic or plausible the perceived spatial impression is. Therefore, a greater magnitude of LEV would not necessarily mean a stronger sense of physical presence. This is supported by the result showing that the

two-channel stereo conditions (0+2+0) were always significantly lower than the surround and 3D conditions for LEV, but not in terms of Presence; 0+2+0 was not significantly different from not only 0+5+0 but also 4+5+0 and 9+10+3 for Brass Band. The present study also provides new evidence that Presence is associated with overall tonal quality (OTQ). Within the context of this study, which is downmixing quality, it is considered that tonal naturalness and realism might have been more important for Presence than LEV, whereas LEV might have dominated OLE perception. Further research is required to define relationships between low-level spatial/timbral attributes and high-level concepts such as presence, involvement and immersiveness [23].

In addition, it is considered that the multiple comparison nature of the listening test may have influenced the results. Multiple comparison is arguably the most popular method of collecting subjects' responses in audio engineering discipline. However, in the quality of experience research community, a single stimulus presentation with absolute categorical rating (ACR) is regarded as a standard response method. Schoffler *et al.* [11] compared the two approaches in their study and found that the results were consistent between them. However, they examined this in terms of OLE only. It may be that ratings on different attributes may depend on the response method. For example, in the present study, the overall tonal quality (OTQ) of the two-channel stereo was mostly rated in the "fair" category whereas the surround and 3D formats were rated in the "good" and "excellent" categories. However, if the two conditions had been tested individually at different times, the stereo version may have been rated in a higher category, or conversely the surround and 3D version may have received a lower rating on the same scale, since the listener would have used his or her absolute judgement on the perceived quality. Stereo is still the format that everyone would be familiar with, and one would not think a high-quality stereo recording is just "fair" in OTQ. It is possible that the stereo conditions were rated to be inferior to the larger formats only because they were compared simultaneously. This raises a question of what would be a more ecologically valid response method for investigating different perceptual attributes as well as OLE.

REFERENCES

- [1] Dolby Laboratories, Inc., "Dolby Atmos," <https://www.dolby.com/technologies/dolby-atmos> (Accessed 6 Nov. 2021).
- [2] J. Herre, J. Hilpert, A. Kuntz and J. Plogsties, "MPEG-H audio—The new standard for universal spatial/3D audio coding," *J. Audio Eng. Soc.*, **62**, 821–830 (2014).
- [3] Sony Europe B.V., "360 Reality Audio," <https://www.sony.co.uk/electronics/360-reality-audio> (Accessed 6 Nov. 2021).
- [4] H. Lee, "Multichannel 3D microphone arrays: A review," *J. Audio Eng. Soc.*, **69**, 5–26 (2021).
- [5] ITU-R, "Advanced sound system for programme production," Recommendation BS.2051-2 (2018).
- [6] K. Hamasaki, T. Nishiguchi, K. Hiyama and K. Ono, "Advanced multichannel audio systems with superior impression of presence and reality," *116th Convention Audio Eng. Soc.*, convention paper 6053 (2004).
- [7] W. Howie, R. King and D. Martin "A three-dimensional orchestral music recording technique, optimized for 22.2 multichannel sound," *141st Convention Audio Eng. Soc.*, convention paper 9612 (2016).
- [8] T. Kamekawa and A. Marui, "Evaluation of recording techniques for three-dimensional audio recordings: Comparison of listening impressions based on difference between listening positions and three recording techniques," *Acoust. Sci. & Tech.*, **41**, 260–268 (2020).
- [9] H. Lee and D. Johnson, "An open-access database of 3D microphone array recordings," *147th Convention Audio Eng. Soc.*, e-Brief 543 (2019).
- [10] A. Silzle, S. George, E. Habets and T. Bachmann, "3D audio quality evaluation: Theory and practice," *Proc. Int. Conf. Spatial Audio*, Erlangen, Germany, pp. 129–138 (2014).
- [11] M. Schoeffler, A. Silzle and J. Herre, "Evaluation of spatial/3D audio: Basic audio quality versus quality of experience," *IEEE J. Sel. Top. Signal Process.*, **11**, 75–88 (2017).
- [12] R. King, B. Leonard and J. Kelly, "Height channel signal level in immersive audio—How much is enough?" *146th Convention Audio Eng. Soc.*, e-Brief 492 (2019).
- [13] ITU-R, "Methods for the subjective assessment of small impairments in audio systems," Recommendation BS.1116-3 (2015).
- [14] D. Cabrera and S. Tilley, "Vertical localization and image size effects in loudspeaker reproduction," *AES 24th Int. Conf. Multichannel Audio—The New Reality*, conference paper 46 (2003).
- [15] H. Lee, "Perceptual band allocation (PBA) for the rendering of vertical image spread with a vertical 2D loudspeaker array," *J. Audio Eng. Soc.*, **64**, 1003–1013 (2016), <https://doi.org/10.17743/jaes.2016.0052>
- [16] J. Vilkamo, A. Kuntz and S. Füg, "Reduction of spectral artifacts in multichannel downmixing with adaptive phase alignment," *J. Audio Eng. Soc.*, **62**, 516–526 (2014).
- [17] C. Gribben and H. Lee, "Towards the development of a universal listening test interface generator in Max," *138th Convention Audio Eng. Soc.*, e-Brief 187 (2015).
- [18] ITU-R, "Methods for the subjective assessment of intermediate quality level of audio systems," Recommendation BS.1534-3 (2015).
- [19] T. Hidaka, L. 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. Audio Eng. Soc.*, **98**, 988–1007 (1995).
- [20] G. Paine, R. Sazdov and K. Stevens, "Perceptual investigation into envelopment, spatial clarity, and engulfment in reproduced multi-channel audio," *AES 31st Int. Conf. New Directions in High Resolution Audio*, conference paper 25 (2007).
- [21] A. Raake and S. Egger, "Quality and quality of experience," in *Quality of Experience: Advanced Concepts, Applications and Methods* (Springer-Verlag, Heidelberg, 2014), pp. 11–33.
- [22] S. Agrawal, A. Simon, S. Bech, K. Bærentsen and S. Forchhammer, "Defining immersion: Literature review and implications for research on immersive audiovisual experiences," *J. Audio Eng. Soc.*, **68**, 404–417 (2020).
- [23] H. Lee, "A conceptual model of immersive experience in extended reality," *PsyArXiv*, <https://doi.org/10.31234/osf.io/sefkh>.
- [24] F. Biocca, "The cyborg's dilemma: Progressive embodiment

- in virtual environments,” *J. Comput.-Mediat. Commun.*, **3**(2) (1997).
- [25] H. Lee and D. Johnson, “3D microphone array comparison: Objective measurements,” *J. Audio Eng. Soc.*, accepted for publication (2021).
- [26] H. Lee and F. Rumsey, “Investigation into the effect of interchannel crosstalk in multichannel microphone technique,” *118th Convention Audio Eng. Soc.*, convention paper 6374 (2005).
- [27] R. Wallis and H. Lee, “Vertical stereophonic localization in the presence of interchannel crosstalk: The analysis of frequency-dependent localization thresholds,” *J. Audio Eng. Soc.*, **64**, 762–770 (2016).
- [28] H. Lee and C. Gribben, “Effect of vertical microphone layer spacing for a 3D microphone array,” *J. Audio Eng. Soc.*, **62**, 870–884 (2014).
- [29] F. Rumsey, S. Zieliński and R. Kassier, “On the relative importance of spatial and timbral fidelities in judgments of degraded multichannel audio quality,” *J. Acoust. Soc. Am.*, **118**, 968–976 (2005).
- [30] J. Thompson, A. Warner and B. Smith, “An active multi-channel downmix enhancement for minimizing spatial and spectral distortions,” *127th Convention Audio Eng. Soc.*, convention paper 7913 (2009).
- [31] C. Faller and P. Schillebeeckx, “Improved ITU and matrix surround downmixing,” *130th Convention Audio Eng. Soc.*, convention paper 8339 (2011).
- [32] J. S. Bradley and G. A. Soulodre, “Objective measures of listener envelopment,” *J. Acoust. Soc. Am.*, **98**, 2590–2597 (1995).

APPENDIX

Table A-1 Summary of the technical details of the recordings used for the experiments.

Label	Music	Venue	Microphone configuration
Brass band	Variation on Laudate Dominum by Edward Gregson. Performed by University of Huddersfield Brass Band. Recorded by Hyunkook Lee.	St.Paul’s Concert Hall, Huddersfield, UK. Approx. 26 m (W) \times 14 m (L) \times 15 m (H). RT60 = 2.1 s.	A PCMA-3D [REF]-based main microphone array. The middle layer, raised at approx. 3 m from the floor, consists of eight cardioid microphones arranged in 1 m \times 1 m square, except for SiL and SiR microphones, each of which was placed at 2 m away from the array base point. No FL and FR for M \pm 060 were utilised. The upper layer was vertically coincident with the middle layer; eight hypercardioids and one cardioid (TpC) pointing directly upwards at the same height as the middle layer. The bottom layer consisted of three cardioids facing away from the ensemble, placed at 30 cm above the floor and directly below FLc, FRc and FC microphones of the middle layer. See [9] for more details.
String quartet	The 1st movement of Dvorak string quartet in G major op.106. Performed by members of Up North Orchestra. Recorded by Hyunkook Lee.		
Choir	Cantata Kaido Tosei by Kiyoshi Nobutoki. Performed by Tokyo University of the Arts Symphony Orchestra and Choir. Recorded by Toru Kamekawa.	Sougakudo Concert Hall, Tokyo, Japan. Approx. 18 m (W) \times 36 m (D) \times 15 m (H). RT60 = approx. 2 s.	A Decca Tree (three omnis, raised at 3.5 m from the floor) and a rear ambience pair (omnis, 7 m from the floor) in an approx. 2 m \times 2 m square, with multiple spot microphones that were mixed to 9+10+3. The middle layer used the five omni microphones for FL, FR, FC, BL and BR, with multiple cardioid spot microphones for woodwinds and a pair of cardioids for the choir (4 m from the floor) mixed to FL, FR, FLc, FRc, SiL and SiR. BC was not utilised. The upper layer was mixed using four cardioids for TpFL and TpFR, and four ambience omnis for TpSiL, TpSiR, TpBL, TpBR, TpBC and TpC. The bottom layer used six cardioids pointing towards the woodwind instruments and timpani.
Orchestra	Mars, The Bringer of War, from The Planets by Gustav Holst. Performed by National Youth Orchestra of Canada. Recorded by will Howie.	Music Multimedia Room at McGill University, Montreal, Canada. 24.4 m (W) \times 18.3 m (L) \times 17 m (H). RT60 = 2.5 s.	The middle layer, raised 3.05 m from the floor, consisted of a Decca Tree (three omnis) for FLc, FRc and FC, and two outrigger omnis for FL and FR, all of which are arranged in one line (3.8 m), with two wide cardioids for BL and BR, one omni for BC, and two cardioids for SiL and SiR, which were placed a few meters behind the Decca Tree. The upper layer was raised 5.1 m to 9.3 m from the floor, and used nine cardioids; TpFL, TpFR, TpSiL, TpSiR, TpBL and TpBR faced sideways, whilst TpFC, TpBC and TpC faced backwards. The bottom layer was at 0.94 m from the floor and placed in the same line as the middle layer; two cardioids for BtFL and BtFR were placed in between FL and FLc, and between FR and FRc, respectively. The BtFC cardioid was directly behind the FC microphone. They were tilted about 45° towards the floor and facing towards the floor. See [7] for more details.