OPEN ACCESS Freely available online

J. Woodcock, W. J. Davies and T. J. Cox, "Influence of Visual Stimuli on Perceptual Attributes of Spatial Audio"
*J. Audio Eng. Soc.*, vol. 67, no. 7/8, pp. 557–567, (2019 July/August.).
DOI: https://doi.org/10.17743/jaes.2019.0019

# Influence of Visual Stimuli on Perceptual Attributes of Spatial Audio

**JAMES WOODCOCK*, WILLIAM J. DAVIES, AND TREVOR J. COX,** *AES Member*
(james-s.woodcock@arup.com)

*Acoustics Research Centre, University of Salford, Salford, M5 4WT, UK*

Reproduced audio is often accompanied with visuals (i.e., television, virtual reality, gaming, and cinema). However, the audio technology for these systems is often researched and evaluated in isolation from the visual component. Previous research indicates that the auditory and visual modalities are not processed separately. For example, visual stimuli can influence ratings of audio quality and vice versa. This paper presents an experiment to investigate the influence of visual stimuli on a set of attributes relevant to the perception of spatial audio. Eighteen participants took part in a paired comparison listening test where they were asked to judge pairs of stimuli rendered to 14-, 5-, and 2-channel systems using 10 perceptual attributes. The stimuli were presented in audio only and audio-visual conditions. The results show a significant and large main effect of the loudspeaker configuration for all the tested attributes other than *overall spectral balance* and *depth of field*. The effect of visual stimuli was found to be small and significant for the attributes *realism, sense of space*, and *spatial clarity*. These results suggest that evaluations of audio-visual technologies aiming to evoke a sense of realism or presence should consider the influence of both the audio and visual modalities.

## 1 INTRODUCTION

There has been an increase of interest in spatial audio in recent years, facilitated in part by advances in object-based audio technology and the proliferation of binaural rendering in virtual reality and 360° applications. This is reflected in research (for example, the S3A [1] and ORPHEUS [2] projects have investigated end-to-end object based pipelines), the standardization of formats and metadata models (e.g., MPEG-H [3] and the Audio Definition Model [4]), and the availability of spatial audio Software Developer Kits [5–7]. Audio systems are typically evaluated by eliciting and rating perceptual attributes [8–12]. Despite the fact that reproduced sound is often experienced alongside a visual component (i.e., television, cinema, virtual reality (VR), and augmented reality (AR)), the perceptual evaluation of audio systems typically does not take into account the influence of visual stimuli. As such it is currently not known what impact the presence of visuals has on the perception of spatial audio.

There are several studies demonstrating that the auditory and visual modalities are not processed independently [13]. The perceived location of an auditory object can be influenced by visual stimuli [14, 15]—commonly referred to as the *ventriloquist effect*—or by the direction of the listener's gaze [16]. Speech perception has been shown to be influenced by visual stimuli via the so called *McGurk effect* [17]; when an utterance of the syllable *ba* is dubbed onto a video of a person uttering the syllable *ga*, most listeners report hearing the syllable *da*. Auditory-visual interaction effects have also been measured physiologically. Shams et al. [18] demonstrated that when a single flash is accompanied by two auditory beeps, two flashes are perceived and the activity in the visual cortex for the illusory flash is similar to that evoked by a real flash.

The presence of visual stimuli has been shown to have an effect on preference and quality ratings for reproduced audio. Iwamiya [19] found that subjective evaluations of quality for both audio and video increased when the two modalities were combined, compared to when presented in isolation. A study into the interaction between audio and visual factors in a home theater system [20, 21] showed that both the audio reproduction method and the screen size influenced the mean rating of the attribute *space*. Hollier et al. [22] found that perceived audio quality increased when the audio is accompanied by visuals. Rumsey et al. [23] investigated the effect of visual stimuli on naïve listener ratings of preference. The presence of visuals was found to have a statistically significant but small effect on listener preference—an interaction effect between the presence of visuals and the type of program material was also found.

* Now at Arup, 6th Floor, 3 Piccadilly Place, Manchester, M1 3BN, UK.

Rojas et al. [24] investigated the effect of auditory stimuli on perceived visual fidelity in stereoscopic 3D and found that when white noise was presented concurrently with an image, ratings of visual fidelity were consistently reduced.

Interactions between the auditory and visual modalities have been studied in the context of music. Iwamiya [19] investigated the effects of matched and mismatched (i.e., videos with the "wrong" audio tracks) audio-visual content on the perception of music. Average ratings of congruence were found to be significantly higher for the matched stimuli than the mismatched stimuli. Platz and Kopiez [25] conducted a meta-analysis of studies into the influence of the visual modality on evaluations of music performance. Aggregating the results from 15 previous studies revealed that the influence of the visual component has a medium effect size (quantified using Cohen's *d*). Hendrickx et al. [26] conducted an experiment on audio-visual coherence during live music performances. In a perceptual evaluation of two different mixes of the same performance, it was found that listeners preferred the mix with audio-visual coherence when a video was presented but a spatially unconstrained mix when no video was presented.

Audio-visual interaction has also been investigated in the context of virtual environments. Doukakis et al. [27] investigated the trade-off in the allocation of computational resources between audio and visual rendering with respect to perceived quality of virtual environments. It was found that the visual component dominated quality ratings initially when the available resources were low, but the difference in importance between the two modalities decreased as the available computational resources increased. Riecke et al. [28] found that auditory stimuli can increase the perception of self-motion and presence in virtual environments.

The presence of visual stimuli has been found to influence the perception of soundscapes. Villon et al. [29] found that visuals of urban settings resulted in more negative ratings of soundscapes. Similarly, Hong and Jong [30] found that presenting images of vegetation significantly increased soundscape preference. Preis et al. [31] found significant differences in self reported comfort between reproduced soundscapes with and without a visual component. Visuals have also been found to have an effect on noise annoyance [32, 33] and the perceived effectiveness of noise barriers [34]. Conversely, Cain et al. [35] found that presenting urban soundscapes with and without visuals had no significant effect on semantic differential ratings of soundscape perception.

Considering the above, it is clear that there is an interaction between the auditory and visual modalities. This interaction can affect lower-level features such as localization [14, 15] and higher-level features such as preference [22, 23] and annoyance [32, 33]. Although presenting audio with a corresponding visual component has been shown to have an effect on preference for different audio reproduction methods, it is currently not known which perceptual attributes of spatial audio this affect.

There have been numerous studies that have aimed to elicit perceptual attributes for reproduced sound. These studies generally result in a non-orthogonal set of attributes describing various timbral (such as clarity and coloration) and spatial (such as envelopment and horizontal width) perceptual attributes [36–41, 9, 10, 42]. A recent study by Francombe et al. [11, 12] systematically determined which perceptual attributes contribute to the listener preference of reproduced audio. The study included a wide range of reproduction methods including systems with height up to and including a 22.2 system. Groups of expert and non-expert listeners took part in an elicitation experiment that resulted in two sets of attributes (one for each of the listener groups) covering a range of timbral and spatial attributes as well as higher-level attributes such as *realism* and *sense of space*. This attribute set is utilized in the present study (see Sec. 2.4).

This paper presents the results of a paired comparison attribute rating experiment that aims to address the research question *"Does the presence of visual stimuli influence perceptual attributes of spatial audio reproduced over loudspeakers?"* The following section describes the design and implementation of an experiment to address this question. Secs. 3 and 4 present the results of the experiment and discuss their implications. Finally, conclusions are presented in Sec. 5.

## 2 METHOD

### 2.1 Participants

Eighteen participants took part in the experiment. All of the participants were experienced in taking part in formal listening experiments and reported normal hearing at the time of the experiment. The mean age of the participants was 29.9 years (standard deviation 8.8). Ten participants were male, five participants were female, and three did not provide this information. Ethical approval for the experiment was obtained from the University of Salford Ethics Committee. The participants were paid £10 for their time.

### 2.2 Apparatus

The experiment took place in the audio booth at the University of Salford. This room consists of 18 Genelec 8030A loudspeakers (only 16 speakers were used in the experiment); 10 speakers are located in the horizontal plane (positioned at azimuths +0°, +30°, +45°, +90°, +135°, ±180°, −30°, −45°, −90°, −135°), 4 at approximately +30° elevation (positioned at azimuths +45°, +135°, −45°, −135°), and 4 at approximately $-30^o$ elevation (positioned at azimuths +45°, +135°, −45°, −135°). The speakers are 1.35 m from the center of the array. The reverberation time of the room is around 0.1 s. The level of the loudspeakers were adjusted to produce the same A-weighted equivalent sound level ($L_{Aeq}$) (± 0.5 dB) for a pink noise signal at the central listening position.

Visuals were reproduced via a 42.5 inch Philips BBDL4330QL Full HD display, mounted on the front wall of the room. The size of the display was selected to conform with the recommendations in ITU-R BT.710-4 [43] for the viewing distance in the audio booth.

(a) Sintel (b) BBC Proms

(c) Turning Forest

Fig. 1. Screenshots of the three program items used in the experiment.

### 2.3 Stimuli

Three different program items were used in the experiment, each of which included accompanying moving visuals that were part of the original program. The program items were:

- Sintel, a CGI action movie. The scene consists of the protagonist fighting a dragon and contains action sounds and non-diegetic music. The scene contains no dialogue. The video contains a number of cuts, and there is a direct relationship between the diegetic sounds and the video. The length of the clip is 15 seconds.
- Footage from the BBC Proms. The scene consists of an orchestral performance. The video pans and cuts to different sections of the orchestra while the audio maintains a fixed perspective. The length of the clip is 16 seconds.
- The Turning Forest, a fixed viewpoint from a VR experience. The scene contains immersive background sound, action sounds, non-diegetic music, and narration. The video is an animated static shot of a forest, and there is no relationship between the diegetic sounds and the video. The length of the clip is 19 seconds.

The scenes were selected to provide a variety of different types of content. However, it should be noted that the selection of stimuli was partly limited by the availability of 3D audio content with accompanying visuals. The Sintel scene provided a stimulus where the audio and visual components were intended to be spatially coherent—each of the diegetic sounds in the scene had a corresponding visual counterpart. The BBC Proms scene provided a stimulus where the audio and visual components were spatially incoherent—the diegetic sounds in the scene had visual counterparts, but the video panned and cut to different parts of the orchestra while the audio remained static. The Turning Forest scene provided a stimulus where the diegetic sounds had no visual counterpart—the visual component in this scene was a shot of a forest where the only movement is falling leaves.

Screenshots of the visuals for each of the program items are shown in Fig. 1. Each of the stimuli were available in 22.2 format, allowing downmixes to multiple different speaker layouts to be created.

Three systems, composed of a subset of the speakers described in Sec. 2.2, were used in the test:

- Fourteen channel system[1]: M+000, M-045, M-090, M-135, M±180, M+135, M+090, M+045, B-045, B+045, U-045, U-135, U+135, U+045
- Five channel system: M+000, M-030, M-135, M+030, M+135
- Two channel system: M-030, M+030

[1] This notation indicates whether the speaker is in the bottom (B), middle (M), or upper (U) layer of the reproduction system along with the azimuth in degrees.

The original 22.2 renders of the stimuli were downmixed using matrix methods to the 14-, 5-, and 2-channel systems using an adaptation of the coefficients presented in [44][2]. The program items were loudness matched using the multichannel extension described in Annex I of ITU-R 1770-2 [45]. For each program item in the test, the stimuli were presented to the participants with and without visuals for each speaker layout meaning that there were a total of six different conditions for each program item (three speaker layouts × two visual conditions).

### 2.4 Test Protocol

The listening test followed a paired comparison method. The method of paired comparisons is a widely used technique to measure the relative differences in the attributes of a set of objects—these attributes may be subjective (i.e., preference, annoyance, taste, smell) or objective (i.e., ranks of sports teams or number of citations to scientific journals). For subjective attributes, the method consists of presenting test participants with pairs of stimuli and asking them to select the stimulus that has more of the specified attributes. For example, participants could be presented with every possible pair of a set of sounds and, for each pair, asked to state which they prefer. Considering the small effect sizes found for the influence of visual stimuli in some of the previous literature (see Sec. 1), the paired comparison method was chosen over magnitude rating as it has been shown to provide a higher discrimination power between stimuli than magnitude rating tasks [46].

A separate paired comparison test was completed for each of the program items described in Sec. 2.3. Each participant completed the three tests in a random order. Participants were presented with every possible pair of the six stimuli within each program item (i.e., the three speaker layouts with and without a screen). Participants were instructed to look at the screen displaying the visuals whenever any audio was playing. After randomizing the order of the stimuli, pairs were presented according to a Ross series to ensure the greatest separation between pairs with common items [47].

For each pair of stimuli, participants were required to indicate relative differences between the pair of stimuli on 10 perceptual attributes. Definitions for the 10 attributes are given in Table 1. The attributes used in this study are informed by a systematic investigation conducted by Francombe et al. [11] into perceptual attributes for a range of audio systems (mono up to 22.2). Two sets of attributes were derived, one for expert listeners and one for non-expert listeners. The attributes selected for use in this study are those that were used at greater than chance frequency in the study reported by Francombe et al. [11] by the expert listener group. The final attribute set consists of lower-level timbral and spatial features (i.e., *spectral balance, depth of field, horizontal width, spatial clarity, spectral clarity, and envelopment*) along with higher-level features (i.e., *realism, spatial openness, sense of space, and spatial naturalness*).

[2] These coefficients are provided in the supporting material.

Table 1. Definitions of attributes used in the listening experiment. The attributes and their definitions are derived from an elicitation experiment described in Francombe et al. [11].

| Attribute | Definition |
|---|---|
| Spectral balance | The magnitude of broad cuts and boosts in the spectrum. |
| Depth of field | Perceived proximity of sources. |
| Horizontal width | 360 degree horizontal width. |
| Spatial openness | How claustrophobic the sound feels. The proximity of the 3D sound field. A sense of air/openness. |
| Spatial clarity | Ease of localization of individual sources. |
| Sense of space | The extent to which you feel you are in the same space in which the music/event was performed. |
| Realism | Overall, how realistic it sounds. |
| Spectral clarity | The ability to distinguish different sources based on their spectral content (timbre). |
| Spatial naturalness | How natural the source position is within the 3D image. |
| Enveloping | How immersed/enveloped you feel in the sound field. |

The participants used the Max/MSP interface shown in Fig. 2 to complete the experiment. This was shown on a separate screen to the display showing the visual component of the stimuli. The participants were allowed to listen to the clips in each pair as many times as they wished. The clips started from the beginning whenever one of the "Play clips" buttons was selected.

### 2.5 Analysis

Generally, the outcome of a paired comparison test is a count matrix $C_{j,k}$. Given a set of $J$ objects indexed by $j$ and $k$:

$$C_{j,k} = \begin{cases} n_{j,k} & j \neq k \\ 0 & j = k \end{cases} \tag{1}$$

where $n_{j,k}$ is the number of times object $j$ is selected over object $k$ in a paired comparison test.

A common way of analyzing paired comparison data is the Bradley-Terry model [48] that models the probability that object $j$ is selected over object $k$.

$$P(j > k|\pi_j, \pi_k) = \frac{\pi_j}{\pi_j + \pi_k} \tag{2}$$

where $\pi_j$ and $\pi_k$ are single figure scores for object $j$ and object $k$ on an interval scale for the attribute being assessed in the paired comparison ratings.

Assuming that $C_{j,k}$ can be used to estimate $P(j > k)$, the scale parameters $\pi_j$ and $\pi_k$ can be estimated using a log-linear model [49]. The outcome of this model is a rank ordering of the stimuli on an interval scale. The *R* package *prefmod* [49] with the extension to the basic Bradley-Terry model that allows ties between pairs of objects was used to calculate $C_{j,k}$ for each perceptual attribute.

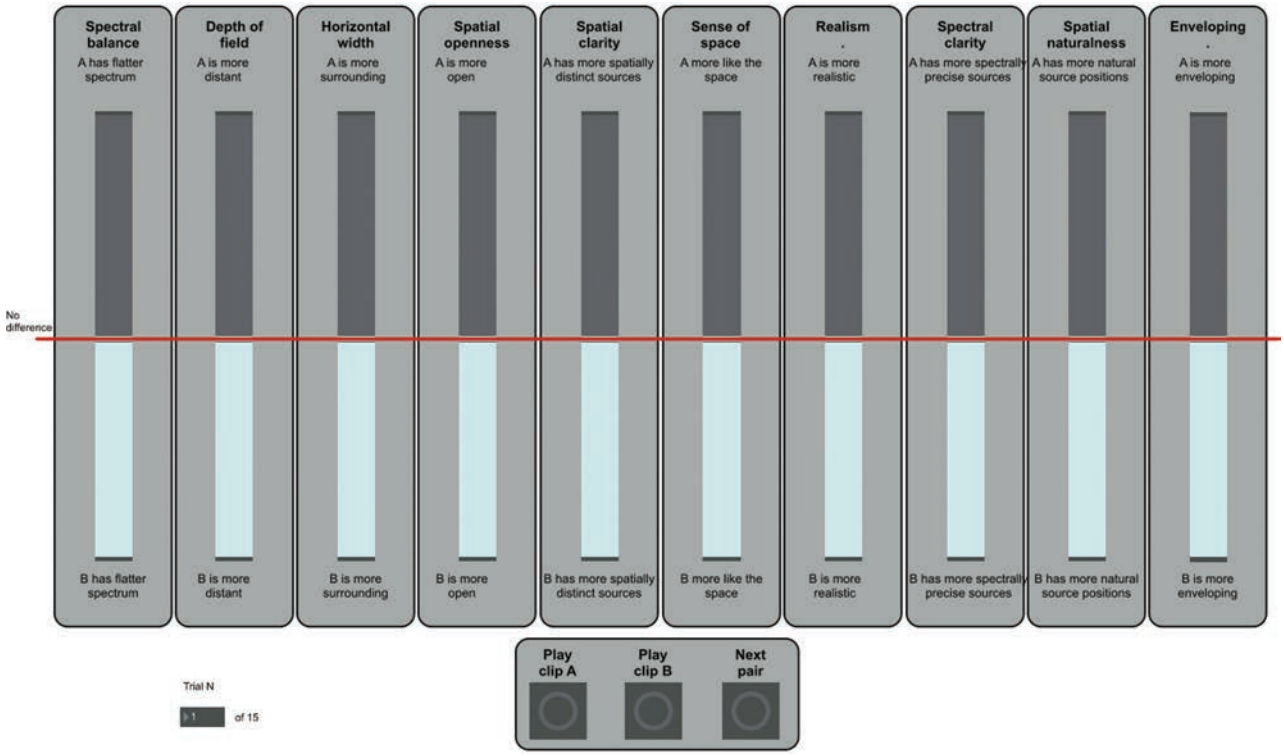



Fig. 2. Interface used in the listening test.

## 3 RESULTS

This section presents the analysis and results from the experiment described in the previous section. In this section CLIP, ATTRIBUTE, and SYSTEM refer to variables that code the different program items, attributes, and systems used in the experiment respectively.

### 3.1 Circular Errors

Intra-participant consistency in paired comparison tests can be assessed using circular error rates [50]. A circular error occurs when a participant makes an inconsistent judgment on a triad of stimuli. For example, an inconsistency would occur if, for a given attribute, a participant were to judge stimulus A > stimulus B, stimulus B > stimulus C, and stimulus C > stimulus A.

Fig. 3 shows the percentage of circular errors for each participant, averaged over all attributes and broken down by CLIP. It can be seen from this figure that participants 3, 15, and 16 generated a higher proportion of circular errors than the other participants. However, they are generally less than 10% suggesting that the ratings made by these participants were still relatively consistent. In a loudness rating experiment, Parizet [50] found that participants with circular error percentages of up to 15% produced comparable results to participants with circular error percentages of less than 1%.

Shapiro-Wilks tests revealed that the circular errors within each level of ATTRIBUTE were not normally distributed. To investigate whether the percentage of circular errors depended on the attribute being rated, a Kruskal-Wallis test was conducted with the circular error percentage as the dependent variable and ATTRIBUTE as the independent variable. This analysis revealed a significant main effect of ATTRIBUTE ($H(9) = 18.6, p < 0.05$) on the circular error percentage, suggesting that the mean circular error percentage differed between attributes. Pairwise Wilcoxon signed rank tests with Bonferroni corrections showed a significantly higher percentage of circular errors for the attribute *depth of field* compared to the attributes *horizontal width, sense of space*, and *spatial openness* ($p_{corrected} < 0.05$). A significant difference in the percentage of circular errors was also observed between the attributes *sense of space* and *realism* ($p_{corrected} < 0.05$).

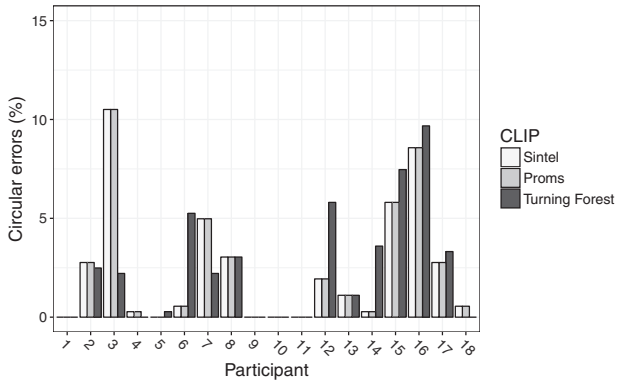



Fig. 3. Percentage of circular errors for each participant, averaged over all attributes.

To investigate whether the program item had a significant effect on the percentage of circular errors, a Kruskal-Wallis test was conducted with circular error as the dependent variable and CLIP as the independent variable. The test

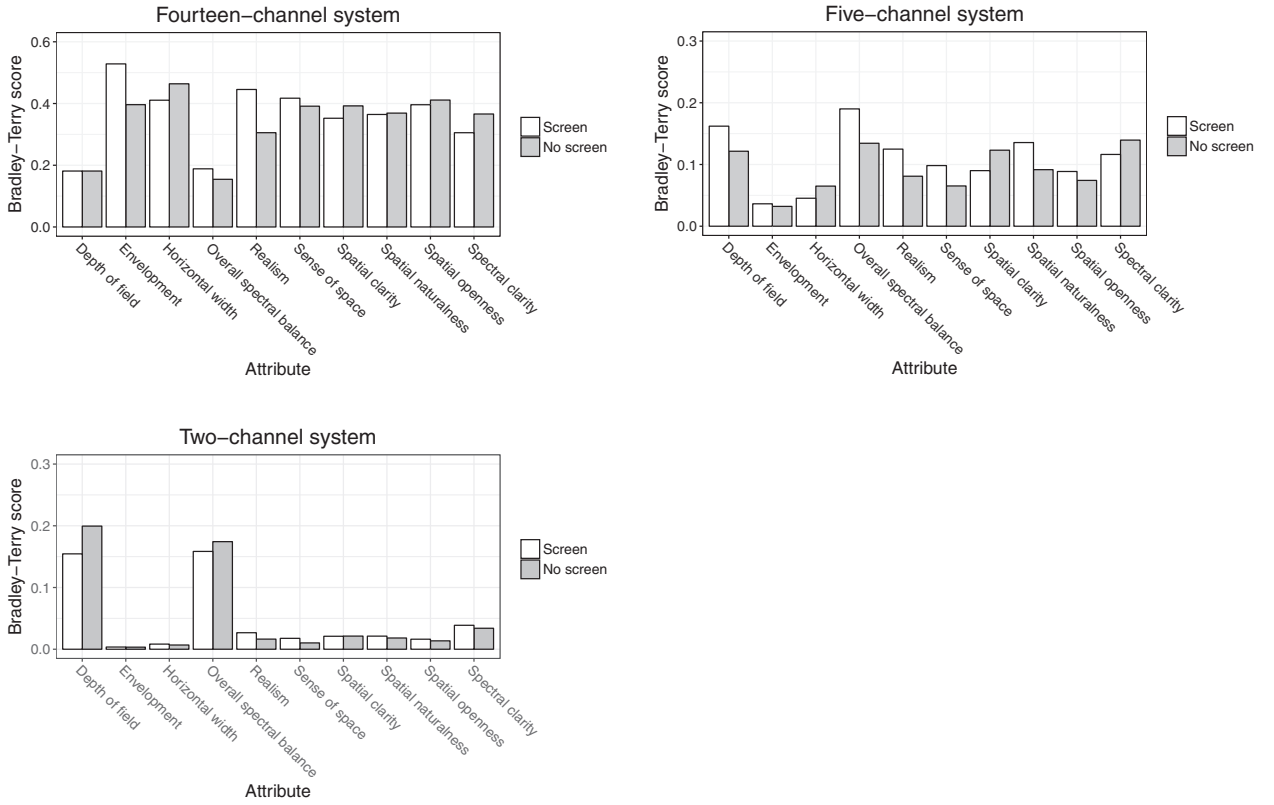



Fig. 4. Bradley-Terry scores for each attribute.

showed that CLIP did not have a significant effect on the percentage of circular errors ($H(2) = 1.82$, $p = 0.40$).

### 3.2 Bradley-Terry Scaling

The paired comparison ratings were converted to attribute scores using the Bradley-Terry model described in Sec. 2.5. Fig. 4 shows the calculated scale values for each system for all of the attributes; these values are calculated over all three program items. In this figure, a higher value means that the system was judged to have more of the given attribute. Within each attribute, the rank order of the systems is generally consistent other than *overall spectral balance* and *depth of field*. As expected, the f14-channel system generally shows the highest value across all attributes, followed by the 5-channel system and the 2-channel system. There is little variation between the different system for the *overall spectral balance* and *depth of field* attributes. This is likely because all of the systems used the same types of loudspeaker positioned equidistant from the listener.

### 3.3 Analysis of Individual Differences

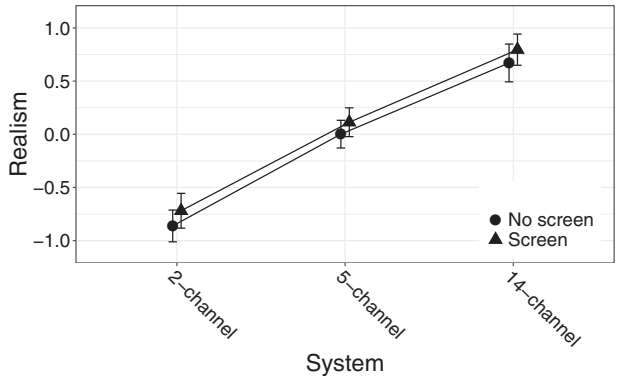
To enable an investigation into where significant differences in the attributes occur, attribute scores were calculated for each individual using the following equation:

$$A_{j,s} = \frac{1}{N_j} \sum_{k \neq j} P_{j,k,s} \tag{3}$$

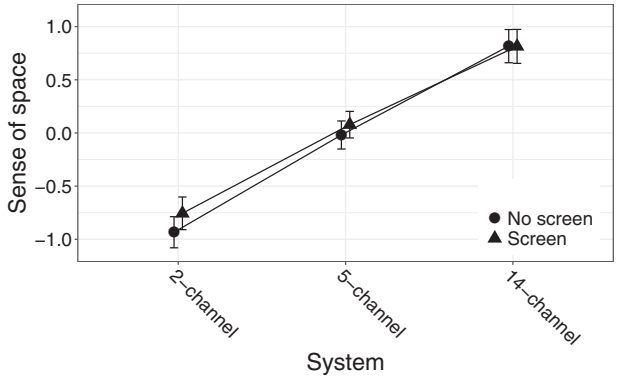
where $A_{j,s}$ is the single figure attribute rating for participant $s$ and stimulus $j$, $N_j$ is the number of times stimulus $j$ was rated by participant $s$, and $P_{j,k,s}$ is the paired comparison attribute rating for stimuli $j$ and $k$ for participant $s$. This method has been utilized previously in the estimation of single figure scores from paired comparison data when individual scores are required [46, 51]. This method of calculating the scale values also has the advantage of taking into account the magnitude of the ratings, whereas the Bradley-Terry scores presented in Sec. 3.2 only take into account which of the stimuli in the pair was selected as having more of a given attribute.

Shapiro-Wilks tests revealed that data in the 78% of conditions met the assumption of normality, therefore parametric statistics are used in the following analysis. For each of the attributes, a repeated measures ANOVA was conducted with within-subject factors of SYSTEM, CLIP, and SCREEN. For all of the attributes apart from *overall spectral balance* and *depth of field* a significant main effect of SYSTEM was found ($p < 0.001$) after Greenhouse-Geisser sphericity corrections were applied. The size of the effect of SYSTEM, judged by the generalized eta-squared ($\eta^2_G$) was large in all cases (mean $\eta^2_G = 0.61$, standard deviation $= 0.08$). A significant main effect of SCREEN was found for the attributes *sense of space* ($F(1, 17) = 4.56$, $p < 0.05$, $\eta^2_G = 0.008$) and *realism* ($F(1, 17) = 6.19$, $p < 0.05$, $\eta^2_G = 0.014$). A significant interaction effect between SCREEN and CLIP was found for the attributes *sense of space* ($F(2, 34) = 3.94$, $p < 0.05$, $\eta^2_G = 0.009$) and *spatial clarity* ($F(2, 34) = 3.95$, $p < 0.05$, $\eta^2_G = 0.01$).

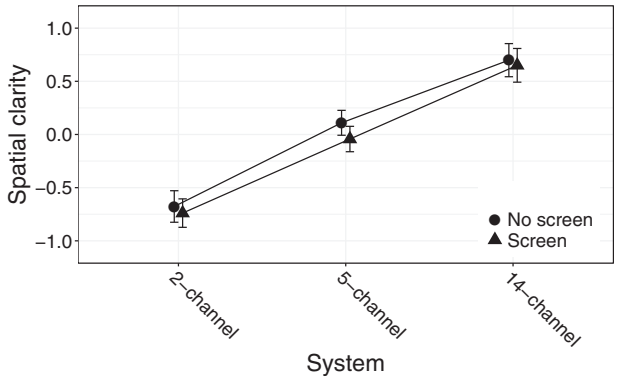
Fig. 5 shows the mean and 95% confidence intervals for the attributes where a significant effect of SCREEN was found. For the *realism* and *sense of space* attributes, it can be seen that when a screen is present there is a small




(a) Realism




(b) Sense of space




(c) Spatial clarity

Fig. 5. Mean and 95% confidence intervals for the attributes where a significant effect of SCREEN was found.

but significant increase in these attributes. For the *spatial clarity* attribute, when the screen is present there is a small but significant decrease in this attribute.

## 4 DISCUSSION

The results presented above answer the research question posed in Sec. 1: *"Does the presence of visual stimuli influence perceptual attributes of spatial audio reproduced over loudspeakers?"* The ANOVA analyses in Sec. 3.3 suggests that the presence of visual stimuli has a significant effect on the perception of spatial audio. This finding is in agreement with the work conducted into the influence of visuals on audio preference and the work investigating the effect of visual stimuli on auditory perception outlined in Sec. 1.

The generalized eta squared ($\eta^2_G$) for the main effect of SCREEN on *sense of space* and *realism* suggests that this effect is relatively small [52] compared to the large main effect of SYSTEM. This finding can be observed in Fig. 5, which shows that the differences between systems is much larger than the differences within systems with and without a screen. The paired comparison methodology was selected because it allows a higher discrimination power than magnitude estimation [46]. However, because participants were able to directly compare the audio-visual and audio only conditions this may have led to the relatively small observed effect sizes. If the experiment were repeated using a single stimulus magnitude rating paradigm, the observed effect sizes may be larger. However, this may be to the detriment of inter- and intra- participant consistency meaning that a greater sample size may be needed.

A significant effect of SCREEN was found for the attributes *realism* and *sense of space* which both relate to the concept of presence. Presence is often used to assess user experience in virtual environments and gaming (e.g., [53]). Lombard and Ditton [54] state that one conceptualization of presence is "the degree to which a medium can produce seemingly accurate representations of objects, events, and people—representations that look, sound, and/or feel like the "real" thing." Another conceptualization of presence is that of transportation whereby the user is transported to another environment or the virtual environment is transported to the user [54]. These two conceptualizations of presence are well described by the attributes *realism* ("*overall, how realistic it sounds*") and *sense of space* ("*the extent to which you feel you are in the same space in which the music/event was performed*"), which were both found to be significantly affected by visual stimuli.

It is interesting to note that the attribute *spatial clarity* was significantly lowered by the presence of the screen. This attribute is defined as the "*ease of localization of individual sources.*" A possible explanation for this is that the screen is constrained to a viewing angle of around 30° whereas the 5- and 14-channel systems are able to reproduce sources 360° around the listener. This means that there is the possibility of an incongruity between the position of

visual sources on the screen and auditory sources that may lead to a degradation in spatial clarity. Future work could investigate the imbalance of audio and visual conditions by investigating the effect of the degree of visual immersion on the perception of audio-visual stimuli.

Understanding the impact of the presence of visual stimuli on the perception of reproduced audio could influence rendering strategies for object-based content. Object-based audio provides the opportunity to optimize rendering based on the target reproduction system, as well as contextual and situational factors regarding the reproduction environment and the listener. Knowledge of the influence of the presence of visual stimuli on the perception of spatial audio could therefore inform the development of rendering strategies that are dependent on whether or not the audio content is accompanied by visuals. Additionally, listener tracking could be used to determine whether or not the listener is currently looking at the screen, and the rendering could be adapted accordingly.

Increasingly, object-based audio renderers are available as part of VR and game development environments [55, 7, 6, 5]. The results presented in this paper suggest that visual stimuli could be used to mask the deficiencies of lower complexity audio rendering algorithms. Further work would be needed using VR reproduction as stimuli to quantify the extent to which this effect could be used. The findings also have implications for object-based audio where knowledge of the reproduction environment can be used to render content in different ways. This means that audio rendering could be adapted to compensate for the decrease in certain perceptual attributes when a screen is not present. Further work would be needed to understand how the parameters of an object-based audio renderer relate to these attributes.

Taken together, these results provide further evidence that multi-modal perceptual information is not processed independently [13]. That the visuals did not effect all of the tested attributes in the same way suggests that the interaction between the auditory and visual modalities is not a simple summation of the two perceptual channels. These findings suggest that audio technologies intended to be used in an audio-visual application should generally be evaluated alongside the intended visual component. However, no significant effect of visual stimuli was found for the lower-level timbral and spatial attributes that were tested. This suggests that if only lower-level spatial and timbral attributes such as spectral balance, envelopment, and clarity are being evaluated, the evaluation could be conducted without visual stimuli.

## 5 CONCLUSION

This study investigated the influence of visual stimuli on a range of perceptual attributes of spatial audio reproduction. Significant main effects were found for the loudspeaker configuration and the presence of visual stimuli. The effect of loudspeaker configuration was found to be large and significant for all of the tested attributes other than *overall spectral balance* and *depth of field*. The effect of visual stimuli was found to be small and significant for the attributes *realism, sense of space*, and *spatial clarity*. The findings suggest that the presence of visuals do not influence lower-level timbral and spatial attributes such as spectral balance and envelopment, but do have an influence on higher-level attributes such as realism and sense of space. This suggests that the influence of visual stimuli may be context dependent and that evaluations of audio-visual technologies aiming to evoke a sense of realism or presence should take into account the influence of both the audio and visual modalities.

## 6 ACKNOWLEDGMENT

The authors would like to thank Lara Harris for running the experiments and the participants. Thank you to Chris Pike from BBC R&D for providing the Turning Forest and Proms clips and Fraunhofer for providing the Sintel clip. This work was supported by the EPSRC program Grant S3A: Future Spatial Audio for an Immersive Listener Experience at Home (EP/L000539/1) and by the BBC Audio Research Partnership.

The experimental data underlying the findings are fully available without restriction, details are available from https://doi.org/10.17866/rd.salford.7932134. Due to copyright restrictions, the stimuli used in the listening experiments is not available from this link.

## 7 REFERENCES

[1] P. Coleman, A. Franck, J. Francombe, Q. Liu, T. de Campos, R. Hughes, D. Menzies, S. Galvez, Y. Tang, J. Woodcock, P. J. B. Jackson, F. Melchoir, C. Pike, F. M. Fazi, T. J. Cox, and A. Hilton, "An Audio-Visual System for Object-Based Audio: From Recording to Listening," *IEEE Transactions on Multimedia*, pp. 1919–1931 (2018 Jan.), doi:http://doi.org/10.1109/TMM.2018.2794780.

[2] M. Weitnauer, C. Baume, A. Silzle, N. Färber, O. Warusfel, N. Epain, T. Herberger, N. Färber, B. Duval, and N. Bogaards, "D2.2: Interim Reference Architecture Specification and Integration Report," Tech. rep., ORPHEUS public deliverable (2017), https://orpheus-audio.eu/public-deliverables/.

[3] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, "MPEG-H 3D Audio—The New Standard for Coding of Immersive Spatial Audio," *IEEE Selected Topics in Signal Processing*, vol. 9, no. 5, pp. 770–779 (2015 Aug.), doi: http://doi.org/10.1109/JSTSP.2015.2411578.

[4] ITU-R rec. BS.2076-1, "Audio Definition Model," ITU-R Broadcasting Service (Sound) Series (2017).

[5] "Google Resonance Audio," https://developers.google.com/resonance-audio/develop/overview, accessed: 2018-06-12.

[6] "Steam Audio," https://valvesoftware.github.io/steam-audio/, accessed: 2018-06-12.

[7] "Unity User Manual (2018.1)/AudioNative Audio Plugin SDK/Audio Spatializer SDK," https://docs.unity3d.com/Manual/AudioSpatializerSDK.html, accessed: 2018-06-12.

[8] N. Zacharov and K. Koivuniemi, "Audio Descriptive Analysis & Mapping of Spatial Sound Displays," presented at the *2001 International Conference on Auditory Display* (2001).

[9] S. Choisel and F. Wickelmaier, "Extraction of Auditory Features and Elicitation of Attributes for the Assessment of Multichannel Reproduced Sound," *J. Audio Eng. Soc.*, vol. 54, pp. 815–826 (2006 Sep.).

[10] S. Choisel and F. Wickelmaier, "Evaluation of Multichannel Reproduced Sound: Scaling Auditory Attributes Underlying Listener Preference," *J. Acoust. Soc. Amer.*, vol. 121, no. 1, pp. 388–400 (2007 Jan.), doi:http://doi.org/10.1121/1.2385043.

[11] J. Francombe, T. Brookes, and R. Mason, "Evaluation of Spatial Audio Reproduction Methods (Part 1): Elicitation of Perceptual Differences," *J. Audio Eng. Soc.*, vol. 65, pp. 198–211 (2017 Mar.), http://doi.org/10.17743/jaes.2016.0070.

[12] J. Francombe, T. Brookes, R. Mason, and J. Woodcock, "Evaluation of Spatial Audio Reproduction Methods (Part 2): Analysis of Listener Preference," *J. Audio Eng. Soc.*, vol. 65, pp. 212–225 (2017 Mar.), doi:http://doi.org/10.17743/jaes.2016.0071.

[13] A. Kohlrausch and S. van de Par, "Audio-Visual Interaction in the Context of Multi-Media Applications," in *Communication Acoustics*, pp. 109–138 (Springer, 2005).

[14] S. Mateeff, J. Hohnsbein, and T. Noack, "Dynamic Visual Capture: Apparent Auditory Motion Induced by a Moving Visual Target," *Perception*, vol. 14, no. 6, pp. 721–727 (1985 Dec.), doi:http://doi.org/10.1068/p140721.

[15] D. Alais and D. Burr, "The Ventriloquist Effect Results from Near-Optimal Bimodal Integration," *Current Biology*, vol. 14, no. 3, pp. 257–262 (2004 Feb.), doi:http://doi.org/10.1016/j.cub.2004.01.029.

[16] J. Lewald, "The Effect of Gaze Eccentricity on Perceived Sound Direction and its Relation to Visual Localization," *Hearing Research*, vol. 115, no. 1–2, pp. 206–216 (1998 Jan.), doi:http://doi.org/10.1016/S0378-5955(97)00190-1.

[17] H. McGurk and J. MacDonald, "Hearing Lips and Seeing Voices," *Nature*, vol. 264, no. 5588, p. 746 (1976 Dec.), doi:http://doi.org/10.1038/264746a0.

[18] L. Shams, Y. Kamitani, S. Thompson, and S. Shimojo, "Sound Alters Visual Evoked Potentials in Humans," *Neuroreport*, vol. 12, no. 17, pp. 3849–3852 (2001 Dec.).

[19] S.-I. Iwamiya, "Interactions between Auditory and Visual Processing when Listening to Music in an Audiovisual Context: 1. Matching 2. Audio Quality." *Psychomusicology: A Journal of Research in Music Cognition*, vol. 13, no. 1-2, p. 133 (1994), doi:http://doi.org/10.1037/h0094098.

[20] W. Woszczyk, S. Bech, and V. Hansen, "Interaction between Audio-Visual Factors in a Home Theater System: Definition of Subjective Attributes," presented at the *99th Convention of the Audio Engineering Society* (1995 Oct.), convention paper 4096.

[21] S. Bech, V. Hansen, and W. Woszczyk, "Interaction between Audio-Visual Factors in a Home Theater System: Experimental Results," presented at the *99th Convention of the Audio Engineering Society* (1995 Oct.), convention paper 4133.

[22] M. P. Hollier and R. Voelcker, "Objective Performance Assessment: Video Quality as an Influence on Audio Perception," presented at the *103rd Convention of the Audio Engineering Society* (1997 Sep.), convention paper 4590.

[23] F. Rumsey, S. Zielinski, R. Kassier, and S. Bech, "Relationships between Experienced Listener Ratings of Multichannel Audio Quality and Naïve Listener Preferences," *J. Acoust. Soc. Amer.*, vol. 117, no. 6, pp. 3832–3840 (2005 May), doi:http://doi.org/10.1121/1.1904305.

[24] D. Rojas, B. Kapralos, A. Hogue, K. Collins, L. Nacke, S. Cristancho, C. Conati, and A. Dubrowski, "The Effect of Sound on Visual Fidelity Perception in Stereoscopic 3-D," *IEEE Transactions on Cybernetics*, vol. 43, no. 6, pp. 1572–1583 (2013 Dec.), doi:http://doi.org/10.1109/TCYB.2013.2269712.

[25] F. Platz and R. Kopiez, "When the Eye Listens: A Meta-Analysis of How Audio-Visual Presentation Enhances the Appreciation of Music Performance," *Music Perception: An Interdisciplinary Journal*, vol. 30, no. 1, pp. 71–83 (2012 Sep.), doi:http://doi.org/10.1525/mp.2012.30.1.71.

[26] E. Hendrickx, J. Palacino, V. Koehl, F. Changenet, E. Corteel, and M. Paquier, "Should Sound and Image Be Coherent during Live Performances?" presented at the *2018 AES International Conference on Spatial Reproduction-Aesthetics and Science* (2018 July), conference paper 11-2.

[27] E. Doukakis, K. Debattista, C. Harvey, T. Bashford-Rogers, and A. Chalmers, "Audiovisual Resource Allocation for Bimodal Virtual Environments," *Computer Graphics Forum*, vol. 37, no. 1, pp. 172–183 (2018 Feb.), doi:http://doi.org/10.1111/cgf.13258.

[28] B. E. Riecke, A. Väljamäe, and J. Schulte-Pelkum, "Moving Sounds Enhance the Visually-Induced Self-Motion Illusion (Circular Vection) in Virtual Reality," *ACM Transactions on Applied Perception (TAP)*, vol. 6, no. 2, p. 7 (2009 Feb.), doi:http://doi.org/10.1145/1498700.1498701.

[29] S. Viollon, C. Lavandier, and C. Drake, "Influence of Visual Setting on Sound Ratings in an Urban Environment," *Applied Acoustics*, vol. 63, no. 5, pp. 493–511 (2002 May), doi:http://doi.org/10.1016/S0003-682X(01)00053-6.

[30] J. Y. Hong and J. Y. Jeon, "Designing Sound and Visual Components for Enhancement of Urban Soundscapes," *J. Acoust. Soc. Amer.*, vol. 134, no. 3, pp. 2026–2036 (2013 Aug.), doi:http://doi.org/10.1121/1.4817924.

[31] A. Preis, J. Kociński, H. Hafke-Dys, and M. Wrzosek, "Audio-Visual Interactions in Environment Assessment," *Science of the Total Environment*, vol. 523, pp. 191–200 (2015 Aug.), doi:http://doi.org/10.1016/j.scitotenv.2015.03.128.

[32] Z. Bangjun, S. Lili, and D. Guoqing, "The Influence of the Visibility of the Source on the Subjective Annoyance due to its Noise," *Applied Acoustics*, vol. 64, no. 12, pp. 1205–1215 (2003 Dec.), doi:http://doi.org/10.1016/S0003-682X(03)00074-4.

[33] L. Maffei, M. Masullo, F. Aletta, and M. Di Gabriele, "The Influence of Visual Characteristics of Barriers on Railway Noise Perception," *Science of the Total Environment*, vol. 445, pp. 41–47 (2013 Feb.), doi:http://doi.org/10.1016/j.scitotenv.2012.12.025.

[34] J. Y. Hong and J. Y. Jeon, "The Effects of Audio–Visual Factors on Perceptions of Environmental Noise Barrier Performance," *Landscape and Urban Planning*, vol. 125, pp. 28–37 (2014 May), doi:http://doi.org/10.1016/j.landurbplan.2014.02.001.

[35] R. Cain, P. Jennings, and J. Poxon, "The Development and Application of the Emotional Dimensions of a Soundscape," *Applied Acoustics*, vol. 74, no. 2, pp. 232–239 (2013 Feb.), doi:http://doi.org/10.1016/j.apacoust.2011.11.006.

[36] T. Letowski, "Sound Quality Assessment: Concepts and Criteria," presented at the *87th Convention of the Audio Engineering Society* (1989 Oct.), convention paper 2825.

[37] F. Rumsey, S. Zieliński, R. Kassier, and S. Bech, "On the Relative Importance of Spatial and Timbral Fidelities in Judgments of Degraded Multichannel Audio Quality," *J. Acoust. Soc. Amer.*, vol. 118, no. 2, pp. 968–976 (2005 Aug.), doi:http://doi.org/10.1121/1.1945368.

[38] N. Zacharov and K. Koivuniemi, "Unravelling the Perception of Spatial Sound Reproduction: Analysis & External Preference Mapping," presented at the *111th Convention of the Audio Engineering Society* (2001 Nov.), convention paper 5423.

[39] N. Zacharov and K. Koivuniemi, "Unravelling the Perception of Spatial Sound Reproduction: Language Development, Verbal Protocol Analysis and Listener Training," presented at the *111th Convention of the Audio Engineering Society* (2001 Nov.), convention paper 5424.

[40] N. Zacharov and K. Koivuniemi, "Unraveling the Perception of Spatial Sound Reproduction: Techniques and Experimental Design," presented at the *AES 19th International Conference: Surround Sound-Techniques, Technology, and Perception* (2001 Jun.), conference paper 1929.

[41] C. Guastavino and B. Katz, "Perceptual Evaluation of Multi-Dimensional Spatial Audio Reproduction," *J. Acoust. Soc. Amer.*, vol. 116, no. 2, pp. 1105–1115 (2004 Aug.), doi:http://doi.org/10.1121/1.1763973.

[42] A. Lindau, V. Erbes, S. Lepa, H. Maempel, F. Brinkman, and S. Weinzierl, "A Spatial Audio Quality Inventory (SAQI)," *Acta Acustica united with Acustica*, vol. 100, no. 5, pp. 984–994 (2014 Sep./Oct.), doi:http://doi.org/10.3813/AAA.918778.

[43] Rec. ITU-R BT.710-4, "Subjective assessment methods for image quality in high-definition television" (1998).

[44] T. Sugimoto, S. Oode, and Y. Nakayama, "Downmixing Method for 22.2 Multichannel Sound Signal in 8K Super Hi-Vision Broadcasting," *J. Audio Eng. Soc.*, vol. 63, pp. 590–599 (2015 Jul./Aug.), doi:http://doi.org/10.17743/jaes.2015.0062.

[45] Rec. ITU-R BS.1770-4, "Algorithms to measure audio programme loudness and true-peak audio level" (2015).

[46] E. Parizet, N. Hamzaoui, and G. Sabatie, "Comparison of Some Listening Test Methods: A Case Study," *Acta Acustica united with Acustica*, vol. 91, no. 2, pp. 356–364 (2005 Mar./Apr.).

[47] H. A. David, *The Method of Paired Comparisons*, vol. 12 (London, 1963).

[48] R. A. Bradley and M. E. Terry, "Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons," *Biometrika*, vol. 39, no. 3/4, pp. 324–345 (1952 Dec.), doi:http://doi.org/10.2307/2334029.

[49] R. Hatzinger, R. Dittrich, et al., "Prefmod: An R Package for Modeling Preferences Based on Paired Comparisons, Rankings, or Ratings," *J. Statistical Software*, vol. 48, no. 10, pp. 1–31 (2012 May).

[50] E. Parizet, "Paired Comparison Listening Tests and Circular Error Rates," *Acta Acustica united with Acustica*, vol. 88, no. 4, pp. 594–598 (2002 Jul./Aug.).

[51] J. Woodcock, A. Moorhouse, and D. Waddington, "A Multidimensional Evaluation of the Perception and Annoyance Caused by Railway Induced Groundborne Vibration," *Acta Acustica united with Acustica*, vol. 100, no. 4, pp. 614–627 (2014 Jul./Aug.), doi:http://doi.org/10.3813/AAA.918741.

[52] R. Bakeman, "Recommended Effect Size Statistics for Repeated Measures Designs," *Behavior Research Methods*, vol. 37, no. 3, pp. 379–384 (2005 Aug.), doi:http://doi.org/10.3758/BF03192707.

[53] J. Takatalo, T. Kawai, J. Kaistinen, G. Nyman, and J. Häkkinen, "User Experience in 3D Stereoscopic Games," *Media Psychology*, vol. 14, no. 4, pp. 387–414 (2011 Dec.), doi:http://doi.org/10.1080/15213269.2011.620538.

[54] M. Lombard and T. Ditton, "At the Heart of it All: The Concept of Presence," *J. Computer-Mediated Comm.*, vol. 3, no. 2 (1997 Sep.), doi:http://doi.org/10.1111/j.1083-6101.1997.tb00072.x.

[55] T. Scudiero, "Graphical Processing Units (GPU)-Accelerated Acoustic Simulation for Interactive Experiences," *J. Acoust. Soc. Amer.*, vol. 141, no. 5, pp. 3455–3455 (2017 Jun.), doi:http://doi.org/10.1121/1.4987165.

## THE AUTHORS

James Woodcock

Bill Davies

Trevor Cox

James Woodcock holds a B.Sc. in audio technology, a M.Sc. by research in product sound quality, and a Ph.D. in the human response to whole body vibration, all from the University of Salford. After receiving his Ph.D. James worked as a research fellow at the University of Salford on the EPSRC funded S3A project investigating topics relating to the perception of complex sound scenes including categorization of audio objects and the evaluation of immersive audio technologies. James currently works as a consultant in the acoustics team at Arup.

•

Bill Davies is professor of acoustics and perception at the University of Salford. He researches human response to complex sound fields in areas such as room acoustics, spatial audio, and urban soundscapes. He led the Positive Soundscape Project, an interdisciplinary effort to develop new ways of evaluating the urban sound environment. Bill also leads work on perception of complex auditory scenes on the S3A project. He edited a special edition of *Applied Acoustics* on soundscapes and sits on ISO TC43/SC1/WG54 producing standards on soundscape assessment. He is an Associate Dean in the School of Computing, Science and Engineering at Salford and a recent vice president of the Institute of Acoustics (the UK professional body). Bill holds a B.Sc. in electroacoustics and a Ph.D. in auditorium acoustics, both from Salford. He is the author of 80 academic publications in journals, conference proceedings, and books.

•

Trevor Cox is professor of acoustic engineering at the University of Salford and a past president of the UK's Institute of Acoustics (IOA). Trevor's diffuser designs can be found in rooms around the world. He is co-author of *Acoustic Absorbers and Diffusers*. He was awarded the IOA's Tyndall Medal in 2004. He is currently working on two major audio projects. Making Sense of Sound is a Big Data project that combines perceptual testing and machine learning. S3A is investigating future technologies for spatial audio in the home. Trevor was given the IOA award for promoting acoustics to the public in 2009. He has presented science shows at the Royal Albert Hall, Purcell Rooms, and Royal Institution. Trevor has presented 25 documentaries for BBC radio including "The Physicist's Guide to the Orchestra." For his popular science book *Sonic Wonderland* (in USA: *The Sound Book*), he won an ASA Science Writing Award in 2015. His second popular science book *Now You're Talking* was published in May 2018. @trevor_cox.