J. Francombe, J. Woodcock, R. J. Hughes, R. Mason, A. Franck, C. Pike, T. Brookes, W. J. Davies, P. J. B. Jackson, T. J. Cox, F. M. Fazi and A. Hilton, "Qualitative Evaluation of Media Device Orchestration for Immersive Spatial Audio Reproduction," *J. Audio Eng. Soc.*, vol. 66, no. 6, pp. 414–429, (2018 June.).

DOI: https://doi.org/10.17743/jaes.2018.0027

Qualitative Evaluation of Media Device Orchestration for Immersive Spatial Audio Reproduction

JON FRANCOMBE,[*1] *AES Associate Member* , **JAMES WOODCOCK**[2], **RICHARD J. HUGHES**[2],
(jon.francombe@bbc.co.uk)
RUSSELL MASON,[1] *AES Member*, **ANDREAS FRANCK,**[3] *AES Member*, **CHRIS PIKE,**[4] *AES Member*,
TIM BROOKES,[1] *AES Member*, **WILLIAM J. DAVIES**[2], **PHILIP J. B. JACKSON**[5],
TREVOR J. COX,[2] *AES Member*, **FILIPPO M. FAZI,**[3] *AES Associate Member*, **AND ADRIAN HILTON**[5]

[1] *Institute of Sound Recording, University of Surrey, Guildford, Surrey, GU2 7XH, UK*
[2] *Acoustics Research Centre, University of Salford, Salford, M5 4WT, UK*
[3] *Institute of Sound and Vibration Research, University of Southampton, Southampton, SO17 1BJ, UK*
[4] *BBC Research and Development, MediaCityUK, Salford, M50 2LH, UK*
[5] *Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, Surrey, GU2 7XH, UK*

The challenge of installing and setting up dedicated spatial audio systems can make it difficult to deliver immersive listening experiences to the general public. However, the proliferation of smart mobile devices and the rise of the Internet of Things mean that there are increasing numbers of connected devices capable of producing audio in the home. "Media device orchestration" (MDO) is the concept of utilizing an ad hoc set of devices to deliver or augment a media experience. In this paper the concept is evaluated by implementing MDO for augmented spatial audio reproduction using object-based audio with semantic metadata. A thematic analysis of positive and negative listener comments about the system revealed three main categories of response: perceptual, technical, and content-dependent aspects. MDO performed particularly well in terms of immersion/envelopment, but the quality of listening experience was partly dependent on loudspeaker quality and listener position. Suggestions for further development based on these categories are given[1].

0 INTRODUCTION

Spatial audio plays an important role in creating and delivering immersive media experiences. The concept of immersive content is multifaceted; the perception of immersion might be created by stimulating multiple senses from all directions, as well as by producing content in which the narrative is engaging and absorbing. In reproduced audio, immersion has generally been achieved by increasing the number of loudspeakers from the ubiquitous two-channel stereo. Loudspeakers can be positioned above, below, in front of, and behind the listener. Systems using from two to twenty-four loudspeakers have been standardized [2], and there has been research into mixing and recording for such formats [3–5]. It is possible to create immersive listening experiences with such systems, but they are challenging to implement in home listening environments (discussed in Sec. 0.1). In this paper an approach to immersive audio reproduction that eschews standardized loudspeaker layouts in favor of utilizing any available sound reproducing devices is introduced. An implementation of this approach is described, and the results from a qualitative evaluation are presented, so that strengths and weaknesses of the proposed approach can be identified.

0.1 Current Methods of Creating Immersive Spatial Audio Experiences

There are three primary methods of representing the sound field to be reproduced over standard loudspeaker arrays. The most common representation is channel-based audio, in which a sound field is represented by a set of loudspeaker signals for a specified layout. The signals may

*To whom correspondence should be addressed. Now at: BBC Research and Development, MediaCityUK, Salford, M50 2LH, UK.

[1] This paper is an extension of the work presented at the Audio Mostly 2017 conference by Francombe et al. [1].

be created in a number of ways; the most common is some variant of amplitude panning, such as vector base amplitude panning (VBAP) [6]. In scene-based audio the sound field is represented as a set of spatial basis functions, most commonly ambisonics where spherical harmonics are used [7]. Finally, in object-based audio, the sound field is represented as a set of audio objects—an object constitutes an audio stream for an individual component of a scene (such as an actor's voice) or a collection of components (such as a choir) with metadata that provide enough information for the renderer to determine how to reproduce the object. The metadata required are determined by the rendering method; however, simple properties such as the object position and level are common to the majority of metadata schemas [8, 9]. The rendering process could theoretically be performed using a number of different algorithms, but often uses VBAP. The difference between channel- and object-based audio is that in the latter, the rendering is delayed until immediately prior to reproduction, enabling easier adaptation to the available loudspeakers, as well as personalization.

There are considerable challenges in creating immersive experiences in realistic domestic listening environments with these representations. In order to achieve an immersive listening experience, many loudspeakers at a range of positions are usually required. When amplitude panning methods are used, the spacing between loudspeakers should be around 60 degrees or less [10] in order to produce virtual sources in the intended directions.

Loudspeakers must also be placed in specified positions. For channel-based transmission, the reproduction format is predefined at the production stage; the channel feeds are transmitted with the expectation that they will be reproduced over the same or a very similar loudspeaker array. The quality of the listening experience is adversely affected when loudspeakers are placed away from the correct, standardized positions [11]. While it is possible to adapt the signals for alternative loudspeaker layouts, methods for doing this have limited flexibility (for example, matrix upmixing [12] or downmixing [13]), or involve significant complexity and a risk of audible artifacts (for example, using source separation or signal analysis and separation techniques [14]). Scene-based audio offers greater flexibility to render to different loudspeaker layouts; however, optimal performance is dependent on having a large number of loudspeakers spaced around a listener [15], with regular sampling on the sphere at a resolution appropriate for the given spherical harmonics series truncation order [16]. Object-based audio removes the limitation of channel-based audio that the loudspeaker layout is predefined, but reproduction is still subject to the limitations of the selected rendering method. Additionally, both VBAP and ambisonic rendering are intended to reproduce a sound field at a defined position in the room. The quality of listener experience is heavily dependent on being in the "sweet spot."

0.2 Evaluation of Spatial Audio Reproduction

Spatial audio reproduction methods are often evaluated by their ability to accurately reproduce the azimuths and elevations of components of the scene. Common criteria include the range of perceived locations that can be reproduced, the accuracy of the perceived location compared to the intended location, or the accuracy of the translation of the scene from production to reproduction [17, 18, 10, 19–24]. For three-dimensional scenes such evaluation methods naturally favor methods with many loudspeakers, which are not feasible in domestic listening environments.

Recent research suggests that rather than aiming for accurate localization or authentic reproduction of the sound field, it may be preferable to optimize other attributes. A preference study conducted by Francombe et al. [25] suggested that envelopment was the most important perceptual factor when comparing different spatial audio reproduction methods. Similarly, Rumsey et al. [26] found that envelopment was more desirable than frontal spatial fidelity. More generally, Mason [27] performed a meta-analysis of audio attribute elicitation studies and found that attributes related to the precise location of a sound were elicited far less frequently than other attributes, such as envelopment, distance, and extent. The definition of envelopment has been widely discussed in the literature [28]. George et al. [29] state that envelopment in multichannel audio "can be created as a result of immersion by a number of direct (dry sources) and indirect (recorded ambience or reverberant content) sound sources present in the reproduction." Francombe et al. [30] elicited a simpler definition: "how immersed/enveloped you feel in the sound field."

0.3 A Proposed Method for Delivering Immersive Spatial Audio Experiences in Domestic Environments

Current reproduction approaches are impractical for widespread uptake of immersive audio over loudspeakers. Therefore, a new approach is needed to enable listeners to access immersive spatial audio listening experiences at home.

Soundbar systems, in which multiple transducers are integrated into a single unit, are a popular way of delivering spatial audio in a domestic environment. However, Walton et al. [31] evaluated two commercially available soundbars and found that listeners preferred a stereo downmix of the five-channel surround sound signals. Binaural techniques are well established for creating spatial audio for headphones [32], but this study focuses on creating shared listening experiences using loudspeakers.

While listeners may not be prepared to install prescribed high channel count systems in their living rooms, it is likely that there are already a number of existing loudspeakers available. These might include traditional discrete loudspeakers (stereo or surround sound systems); wireless audio devices utilizing Wi-Fi or Bluetooth connections; televisions with built-in speakers; soundbars; personal devices such as mobile phones, tablets, laptops, and smart watches; smart assistants; toys; games consoles; and various other domestic appliances. Furthermore, it is increasingly common for such devices to be connected to a data network.

Therefore, they could theoretically be accessed and used as part of an ad hoc spatial audio system. This might comprise a large number of loudspeakers in a range of spatial positions (including different distances and heights as well as azimuths) and, if used intelligently, might be able to provide significant immersion. This integration of a range of devices is referred to here as media device orchestration (MDO), and may ultimately form part of a wider integration of connected devices such as video screens and lighting. The concept of device orchestration is widely used in the Internet of Things field to descibe communication between devices over a network to enable them to work together. Potential use-cases and supporting technology are presented by MPEG [33].

Making optimal use of an ad hoc loudspeaker array is likely to require a variety of rendering methods. An MDO audio system is likely to vary between rooms and even from day to day within the same room as portable devices are moved. Consequently, use of such a system relies on content that can adapt to the devices that are available. This is made possible by using an object-based audio format. The metadata available in existing systems [9, 34, 35] must be extended to include relevant semantic metadata (as discussed in Sec. 1.2), to allow development of a sophisticated rule set for optimal rendering regardless of the available devices.

Such a system is unlikely to exactly reproduce the sound field in the domestic environment as it was created by the producer. However, the optimal listening experience may be created by optimizing high-level perceptual attributes (such as envelopment) rather than maintaining accurate positions. A non-standard array of loudspeakers (including loudspeakers at a range of distances) may also enhance the ability of the system to reproduce distance cues. These are often overlooked in loudspeaker systems designed to have all devices on a sphere with a central listening position.

0.4 Experiment Aims and Paper Outline

The MDO concept represents a significant paradigm shift from current thinking on spatial audio reproduction. There are many technical challenges that must be solved before this could be made widely available. However, in order to validate the concept, it is first necessary to determine the effect that MDO has on listener experience. Having access to loudspeakers in a range of spatial positions might offer the possibility of increased immersion, regardless of the different qualities of the devices. There may also be other benefits and drawbacks of the MDO approach. In this paper the following research questions are addressed: (i) what are the positive aspects of MDO reproduction; and (ii) what are the negative aspects of MDO reproduction? If there are clear positive aspects then this will validate the MDO approach. Determining the negative aspects will highlight areas for further research and development.

An implementation of MDO was developed using object-based audio with an ad hoc reproduction system comprising devices including fixed and portable loudspeakers. This implementation is described in Sec. 1. In Sec. 2 a qualitative evaluation of the system is presented. The evaluation was designed to address the research questions outlined above by collecting positive and negative comments from a panel of listeners and performing thematic analysis to identify the salient perceptual features. The results are discussed in Sec. 3 and an outlook for future research is presented. The findings of the paper are summarized in Sec. 4.

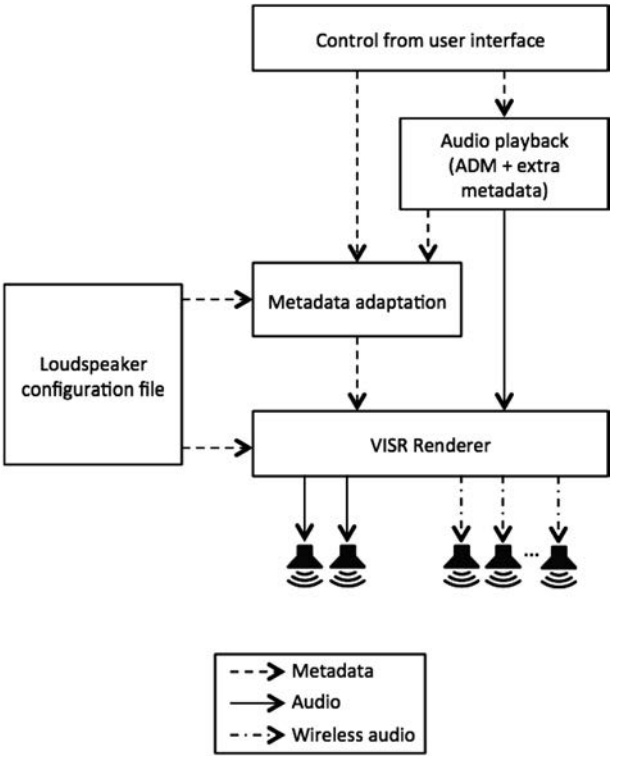


Fig. 1. Diagram of MDO implementation

1 IMPLEMENTATION OF MEDIA DEVICE ORCHESTRATION

In order to investigate the idea that an immersive spatial audio experience could be delivered by augmenting a low channel count reproduction system with an ad hoc collection of connected devices, a demonstration system was established. The system is based on a framework for object-based audio reproduction developed in the S3A project[2], and makes use of the Versatile Interactive Scene Renderer (VISR) [36]. The VISR implements a number of rendering methods, and real-time metadata adaptation can be used to determine the most appropriate method to use for each object. In this case, some objects were rendered using VBAP to a stereo pair of loudspeakers and the remaining objects were rendered to ad hoc devices using direct object-to-loudspeaker routing (DOTLR).

A diagram of the MDO system is shown in Fig. 1. The system relies on metadata that describe properties of the available loudspeakers (Sec. 1.1) and audio objects (Sec. 1.2). The rule set used to determine the rendering method for each audio object is discussed in Sec. 1.3. Finally, the

[2] www.s3a-spatialaudio.org

Table 1. Loudspeaker metadata model

Field	Subfield	Units	Values	Description
ID		–	0–inf	Unique (integer) loudspeaker identifier
Channel		–	0–inf	Physical output channel number
Position				Loudspeaker position relative to central listening position
	Azimuth	Deg.	0–360	
	Elevation	Deg.	0–360	
	Distance	m	0–inf	
Gain		dB	−inf to inf	
Delay		s	0–inf	
Label		–	E.g., "*Main left,*" "*Front table*"	Loudspeaker label, used for display
Auxiliary loudspeaker		–	*False*, *True*	Determines whether the loudspeaker should be considered as part of the main array or as an extra loudspeaker
Quality		–	*Low*, *Medium*, *High*	Loudspeaker quality tag
Function		–	*Primary*, *Secondary*	Used in combination with the audio object function field to control the placement of certain types of sound

user interface that enables control of the system is described in Sec. 1.4.

1.1 Loudspeaker Metadata Format

In the MDO implementation presented in this paper, extra devices are used to augment a low channel count system. This leads to the distinction between the main loudspeaker array (for example a hi-fi system or loudspeakers built into a television) and a set of auxiliary loudspeakers (potentially any sound-emitting device, with a particular focus on personal devices such as mobile phones, tablets, and so on). However, it is also possible to envisage an MDO system with no main set of loudspeakers.

In order to reliably test the MDO concept, the system eschewed wireless communication in favor of wired analog audio connections. The system utilized a high quality stereo pair of studio loudspeakers (Genelec 8030A), augmented by four auxiliary loudspeakers—small Bluetooth-enabled consumer speakers (three Sony SRS-X11s and one B&O Beoplay A2).

In order for the system to make appropriate choices about how to route the audio objects, it was necessary to manually create additional metadata to describe the available loudspeakers (i.e., more than the physical positioning information required for VBAP). The metadata model used to describe the loudspeakers is shown in Table 1.

1.2 Audio Object Metadata Format

The audio content was stored as broadcast wave (BW64) files [37] containing audio definition model (ADM) metadata [9] as described by ITU-R rec. BS.2388-1 [38]. Extra metadata were added to facilitate the rendering method selection and choice of loudspeaker routing for DOTLR. This was added into an additional XML data chunk in the broadcast wave header. The metadata model is detailed in Table 2. It comprises basic metadata stored within the ADM standard and additional time-invariant metadata added to facilitate MDO.

1.3 Metadata Adaptation and Rendering

In this implementation of MDO, scenes are rendered through a combination of VBAP and DOTLR, facilitated using metadata adaptation and object-based rendering. The *VISR* software framework provides flexible rendering of multiple object types, including point source and plane wave objects, but also channel objects that are routed to a specific loudspeaker designated by a channel ID. All objects in the original scenes are either point or plane objects. MDO was performed by processing the metadata for each object and selectively transforming certain objects into channel objects using a simple rule set (applied automatically and in real time). The rule set is implemented in the *Metadapter*, a Python software framework for flexible and extensible adaptation of metadata.

When the MDO processing is turned off, all objects are rendered to stereo using VBAP. When the processing is turned on, the rule set described below uses the metadata to determine a set of suitable loudspeakers for each object.

- If the *Force into auxiliary* flag is set, then only loudspeakers with the *Auxiliary loudspeaker* flag set to *True* can be selected. This flag enabled creative decisions to be made when producing the audio object metadata, i.e., allowing specific objects to be deliberately removed from the main speakers even if their locations were within the range of the stereo pair. The high quality stereo loudspeakers were not included in the set of auxiliary loudspeakers.
- Only loudspeakers tagged at the same *Quality* as the audio object can be selected. This ensures, for example, that audio objects with a high amount of low-frequency energy are not played from small, low-quality devices. If the *Quality* of the audio object is set to *Any*, then any loudspeaker can be selected.
- If the *Function* is set to *Narrator*, then only a loudspeaker tagged as *Primary* can be selected. If the *Function* is set to *Ambience*, then only a loudspeaker tagged as *Secondary* can be selected.

Table 2. Audio object metadata model

Field	Subfield	Units	Values	Description
ID		–	0–inf	Unique audio object identifier
Channel		–	0–inf	Renderer input channel on which the audio content for an object is received
Type			*Plane*, *Point*, *ChannelObject*	Object type flag. The VBAP rendering used for *Plane* and *Point* objects does not differentiate between the object types or account for the distance; it simply renders to a given direction.
	Azimuth	Deg.	0–360	(*Plane* objects only)
	Elevation	Deg.	0–360	
	Distance	m	0–inf	
	X	m	0–inf	(*Point* objects only)
	Y	m	0–inf	
	Z	m	0–inf	
	Output channel	-	0–inf	The loudspeaker ID to which a ChannelObject will be routed (ChannelObject objects only)
Level		dB	−inf to inf	Gain applied to an object
Label		–	E.g., "*Narrator,*" "*Water sounds*"	Audio object label, used for display
Force into auxiliary		–	*False*, *True*	If this flag is set to *True*, the object will be forced into an auxiliary loudspeaker if there are any suitable loudspeaker available (i.e., those that conform to any specified quality and function requirements)
Target loudspeaker quality		–	*Low*, *Medium*, *High*, *Any*	Defines a loudspeaker quality that must be used if the object is routed to an auxiliary loudspeaker
Function		–	*Narrator*, *Ambience*, *Any*	Used in combination with the loudspeaker *Function* field to control the placement of certain types of object

If suitable loudspeakers are found, the *Type* of the current object is changed to *Channel object*, and the *ID* of the loudspeaker closest to the object's original position (i.e., with the smallest Euclidean distance) is assigned. If no suitable loudspeakers are found, the object *Type* is not changed (and consequently the object is rendered using VBAP to the stereo bed).

The potential for extending this rule set is discussed in Sec. 3.

1.4 User Interface

The user interface for the demonstration system is shown in Fig. 2.

The interface enabled switching between stereo and MDO reproduction, selection of program material, control of overall level, and transport control of playback. Each loudspeaker was individually visualized and could be enabled or disabled; the reproduction would adapt in real time to the available devices. The labels or IDs of objects being routed to each loudspeaker were displayed. The user interface used open sound control (OSC) messages to communicate changes to the *Metadapter*.

2 EVALUATION OF MDO IMPLEMENTATION

The MDO implementation described in Sec. 2 was set up in an ITU-R BS.1116 [39] listening room at the University of Salford. Twenty participants experienced a demonstration in three groups. A questionnaire was used to collect qualitative information about the listening experience[3]. In the following sections the demonstration setup is detailed (Sec. 2.1) and the data collection procedure outlined (Sec. 2.2). The results from the questionnaire are presented in Sec. 2.3 and summarized in Sec. 2.4.

2.1 Demonstration Setup

In Fig. 3, the loudspeaker layout used for the demonstrations is shown. The left, rear, and right smaller Bluetooth-enabled speakers (Sony SRS-X11s) were located on a chair arm, low shelf, and high shelf at heights of approximately 0.5 m, 1.0 m, and 1.6 m respectively. The higher quality Bluetooth-enabled speaker (B&O Beoplay A2) was positioned on a coffee table at a height of approximately 0.4 m. These positions were selected as they are representative of possible positions in a real living room.

The loudspeakers were approximately level-aligned by reproducing a pink noise signal from each and adjusting to produce approximately the same loudness (determined by ear) at the central listening position. The loudspeakers were

[3] The data described in this section can be accessed at https://doi.org/10.17866/rd.salford.5589856.

Fig. 2. User interface for MDO demonstration

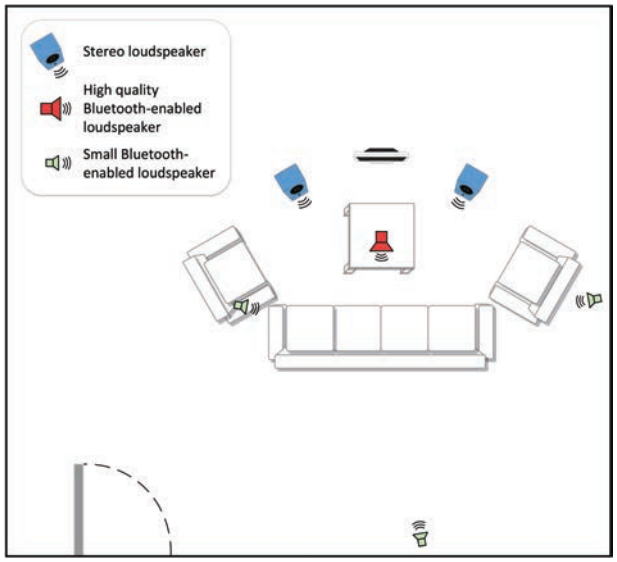


Fig. 3. Loudspeaker positions used in the demonstration (drawn to scale)

also approximately time-aligned at this position[4] by reproducing clicks from each pair of loudspeakers and adjusting a variable delay until there was no audible difference in arrival time. The calibration gains and delays were defined in the loudspeaker metadata (Sec. 1.1).

The participants were played three content items covering a range of genres from audio-only broadcast content.

1. *The Turning Forest*: an object-based audio drama scene [40].
2. *Just Another Frame* by the Hotel Whisky Foxtrot: an object-based pop track, originally mixed in a 22-channel system.
3. A radio advert (originally produced in stereo; remixed using object-based audio in a 22-channel system).

Each program item contained multiple audio objects that could be routed to the additional MDO speakers according to the metadata adaptation and rendering described in Sec. 1.3. Considering this, the responses to the survey questions detailed in Sec. 2.2 could be influenced by the MDO system, the object routing rules, or the program item.

The reproduction was switched (by the demonstration leader) between stereo and MDO rendering multiple times throughout the demonstration to allow the participants to compare the differences between the two reproduction methods. Additionally, the interface was used to enable or disable individual auxiliary loudspeakers to demonstrate the real-time adaptation performed by the system.

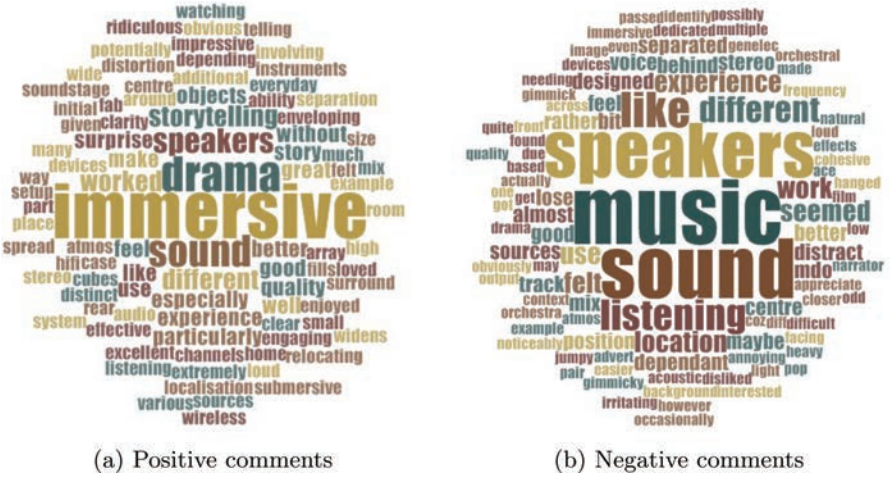
2.2 Data Collection

Immediately following the demonstration, participants were asked to complete a questionnaire featuring three main questions.

1. What did you like/what were the good things about the media device orchestration system?
2. What didn't you like/what were the bad things about the media device orchestration system?
3. [Do you have] any other general thoughts?

Responses were collected as free text data. Twenty participants completed the questionnaire. The respondents were undergraduate and masters level students at the University of Salford; eighteen participants reported that they had some experience of working with audio in a professional capacity. Sixteen of the twenty participants stated that they had professional experience in audio engineering. Consequently, the listeners are likely to be skilled in articulating the perceptual features and attributes of the systems

[4] Some consumer loudspeakers introduce a small delay, so time-alignment was necessary even without wireless transmission.



(a) Positive comments (b) Negative comments

Fig. 4. Word clouds indicating frequency of word use in responses to the questions "*What did you like/what were the good things about the media device orchestration system?*" (left pane) and "*What didn't you like/what were the bad things about the media device orchestration system?*" (right pane)

under investigation. However, the results are potentially less generalizable to a wider population. As the purpose of the study is to understand the positive and negative aspects of the MDO concept, rather than to conduct a broad hedonic evaluation, experienced listeners were preferred to naive listeners.

2.3 Analysis

As an initial analysis of the free text data, word clouds were generated for the responses to questions one and two (see Figs. 4a and 4b). These figures indicate the frequency of word usage in the responses to the two questions. The size of each word is proportional to the number of times it was used.

The figures were generated after removal of stop words and stemming the words so that, for example, the words "listen," "listened," and "listening" would have the same stem and be counted as the same word. The word clouds were generated using *NVivo 11*. From Fig. 4a it can be seen that the most commonly used word in the responses to positive aspects of the MDO demo was "immersive"; also, the specific content type "drama" appears frequently in the responses. The word cloud related to the negative responses shown in Fig. 4b indicates frequent mention of the specific content type "music," and the terms "speakers" and "sound." Unlike the positive terms, there is no prominent adjectival word.

Although the word clouds shown in Figs. 4a and 4b give an initial insight into the frequency of word use in the response to the different survey questions, they do not provide any context around how these words were used. Therefore, a more detailed analysis of the open text data was conducted using thematic analysis [41]. Thematic analysis is a qualitative method that aims to identify themes or patterns in a set of data. This is done through a process of coding salient features of the data in a systematic fashion followed by a collation of the resulting codes into themes.

An inductive approach was used, with the identified codes and themes being driven by the data. Although every effort was made to ensure the analysis was data-driven, it should be acknowledged that researchers cannot completely free themselves from theoretical or epistemological preconceptions; this shortcoming is common to all types of qualitative data analysis [41].

Braun and Clarke [41] outline the main stages of thematic analysis.

1. Familiarization with the data set
2. Generation of initial codes
3. Searching for themes
4. Reviewing themes
5. Defining/naming themes

This process is performed iteratively until no new codes or themes emerge. In the context of thematic analysis, a code is a grouping of related ideas in the data (examples of codes generated in the present study include "listener position" and "quality/type of loudspeakers") and a theme is a collection of related codes (an example of a theme in the present study is "physical setup").

The thematic anaysis was conducted as a group exercise by three of the paper's authors. The 60 responses (20 positive, negative, and general responses) were split into 110 items that each expressed a single idea. From these data, 31 codes were generated. Fig. 5 shows the frequency of usage for each of the codes broken down by whether the coded data appeared in the positive, negative, or general comments section of the survey.

Following this initial coding of the data, the codes were grouped into related themes. This process was repeated a number of times often resulting in related themes being merged. From the raw codes 13 themes were generated; these are listed with definitions and examples in Table 3. The relationships between the raw codes and concepts are shown in the dendrogram in Fig. 6. In this figure the

Table 3. Definitions of the concepts generated in the thematic analysis. Text color in the "example response" column indicates positive (green) or negative (red) responses.

Theme	Definition	Example responses
Spatial attributes	Spatial attributes of the reproduction	"*Was very enveloping*"; "*Liked the spread of sound*"
Clear sounds	Clarity of sounds in the reproduction	"*Sources were clear and distinct*"; "*Sources too separated*"
Cohesion	Cohesion of the overall reproduction	"*Didn't sound like a cohesive reproduction.*"; "*[Different] sounds obviously positioned in space.*"
Loudness balance	Relative balance of sounds in the reproduction	"*Some of the sounds behind were a bit too loud*"; "*Some of the smaller speaker sounds were lost*"
Timbre	Timbral aspects of the reproduced sound scene	"*An unnatural timbre*"
Cognition and evaluation	Relating to understanding, hedonic evaluation, and emotional response to the reproduction	"*Worked very well for storytelling*"; "*Gimmicky effects of voice behind are distracting rather than immersive.*"
Listening mode	How the reproduction is listened to (i.e., background music vs attentive listening)	"*Wouldn't quite work for background music, but for dedicated listening would be good.*"
Physical setup	Height, position, proximity, and type of loudspeakers used in the MDO setup	"*Found the closer speakers annoying*"; "*It's very dependent on the location on where you are sitting.*"
Practicality	How practical the system is to set up	"*Uses everyday devices that are potentially wireless*"; "*Needing multiple devices + speakers*"
Rendering	How the MDO content was rendered	"*Having different objects on different speakers*"; "*Occasionally when a sound passed from one speaker it sounded a bit jumpy.*"
Effect on program type	Effect of MDO on different types of content	"*The immersive feeling of the drama*"; "*Worked better with drama*"
Effect on object type	Effect of MDO on different types of audio object	"*Loved the atmos*"; "*Enjoyed FX*"
Audio-visual	Effect of MDO on reproduction including visuals	"*Narrator voice seemed to distract from the screen*"

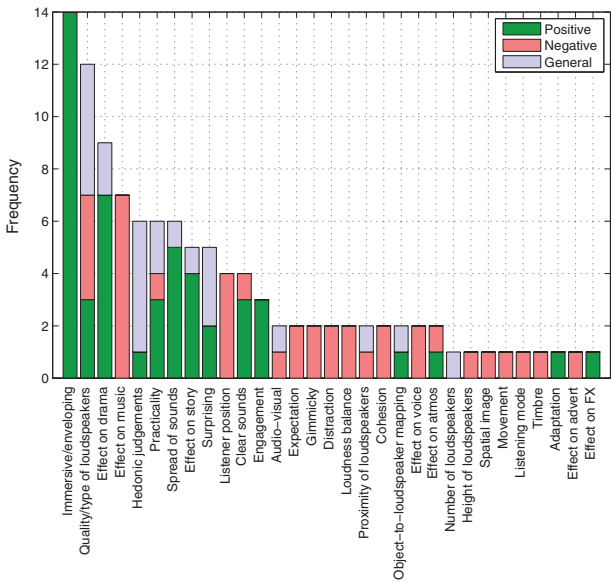


Fig. 5. Frequency of code use (positive, negative, and general responses)

numbers below the labels represent the number of responses underlying that theme for positive, negative, or general comments respectively. Three high-level themes—content, technical, and perceptual—were generated from the concepts.

2.4 Summary of Results

The analysis reported in this section aimed to gather information on the positive and negative aspects of MDO. From the thematic analysis presented in Sec. 2.3, it was found that the responses to the questionnaire could be grouped at the highest level into three categories: content, technical, and perceptual. The frequency of codes associated with these high-level themes suggests that MDO had a strong positive effect on perceptual aspects (32 positive codes compared to 14 negative and 10 general), a tendency towards a negative effect on technical aspects (12 negative codes compared to 8 positive and 10 general), and a tendency towards a negative effect on content related aspects (12 negative codes compared to 9 positive and 3 general). The frequency of positive comments suggests that MDO has a positive effect on listener experience and therefore has potential for creating immersive listening experiences. The results are discussed in more detail in Sec. 3.

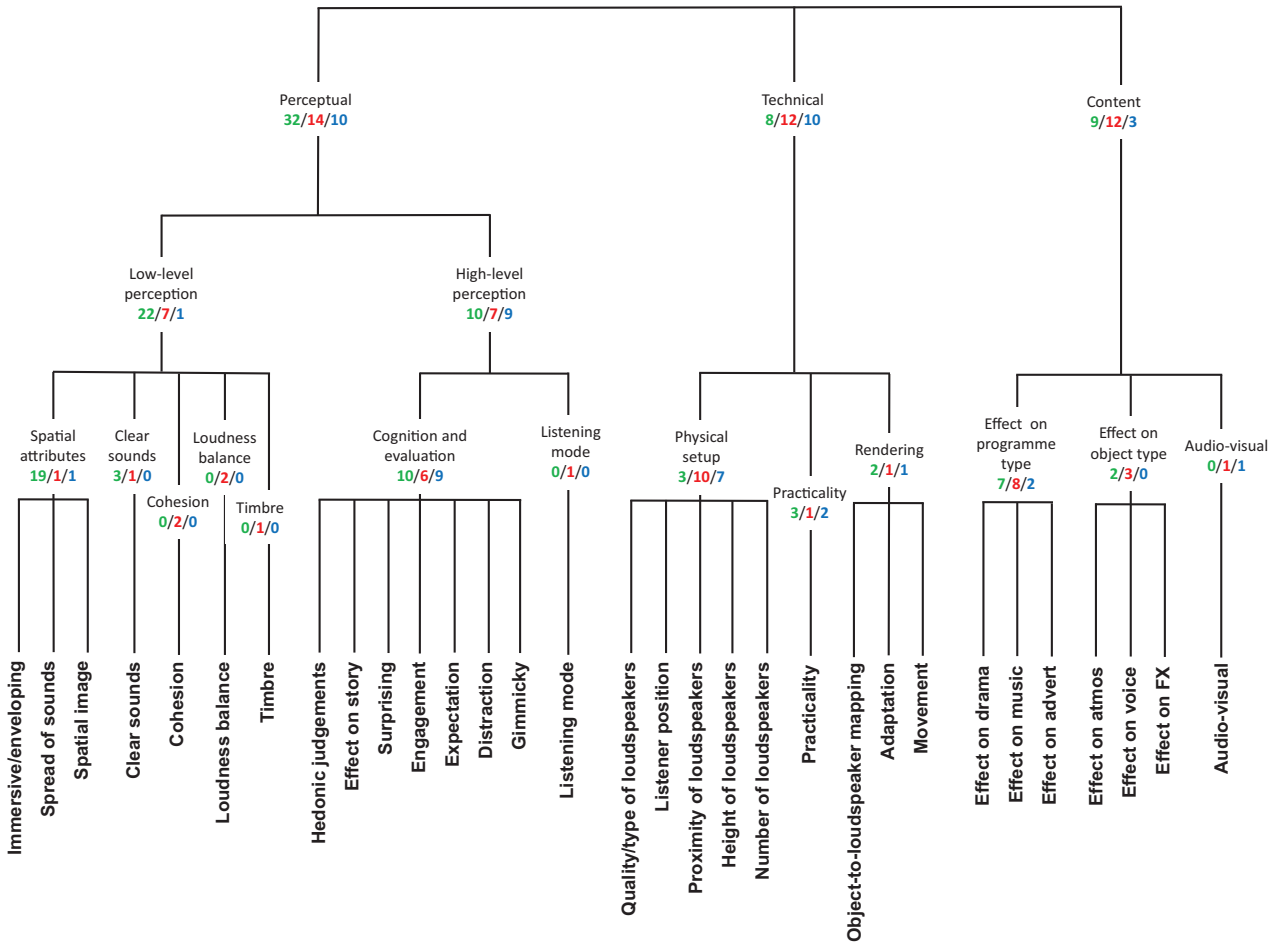


Fig. 6. Dendrogram showing groupings of codes generated in the thematic analysis. Numbers indicate the total frequency of responses in each category broken down into positive/negative/general comments. At each level of the dendrogram the codes and themes are sorted from left to right in decreasing order of frequency.

3 DISCUSSION AND SUGGESTIONS FOR FUTURE WORK

The aim of the experiment reported above was to determine the positive and negative effects of MDO on listener experience so that this approach to spatial audio reproduction could be validated and areas requiring further research and development could be identified. The analysis presented in Sec. 2.3 suggested that MDO is potentially beneficial; the benefits are further highlighted in the detailed analysis below. However, the analysis also highlighted a number of current weaknesses. Further research is required to determine how to best exploit and enhance aspects relating to perceived positive traits, while improving on the negative areas. The rich qualitative data set presented above provides specific areas where further work could be of most use.

The discussion in this section considers the positive and negative effects of MDO and is grouped into three main topics based on the categories found in the analysis: *perceptual* (considering understanding of the listener experience and evaluation through formal scientific comparisons with other reproduction methods); *technical* (considering the implementation challenges and how best to deliver the experience); and *content* (considering how the experience is created and the effect of program type).

3.1 Perceptual

MDO was shown to evoke changes in a number of low- and high-level perceptual attributes that made up the perceptual category. This category had the largest number of positive comments (32 positive, 14 negative, and 10 general comments). A large positive effect on low-level perception was due to the spatial attributes concept, which grouped the codes immersive/enveloping, spread of sounds, and spatial image. In particular, the immersive/enveloping[5] code received 14 positive comments (and no negatives)—twice as many as any other code—and was mentioned alongside the spread of sounds (which had the third highest frequency of positive comments, $N = 5$) as well as effect on drama, effect

[5] From the underlying data, it is clear that this refers to the percept of being immersed or enveloped in a soundfield rather than the higher-level perception of being immersed in the narrative of the content. As the word stem *immers-* was the most commonly used in the data to refer to this percept, it will be used in this context throughout the remainder of this paper.

on story, and engagement. In a study into the relationship between listener preference and perceptual attributes for a wide range of spatial audio systems (mono to 22-channel), Francombe et al. [25] found that the attribute "envelopment" has the largest influence on listener preference. This suggests that MDO could provide significant improvements to the listener experience (compared to stereo) by increasing immersion or envelopment. This finding could be generalized to more standard reproduction methods; for example, increasing the tolerance in loudspeaker positions or using lower-quality speakers for the rear or height channels in surround sound systems.

Comments such as "[feeling] center of the story, part of the experience, not watching it" and "much more immersive than stereo" suggest that MDO could be exploited to create immersive content, with opportunities to investigate how to best deliver these experiences. The spatial attributes concept had just one negative comment, in the spatial image code; the loss of stereo image was found to be "slightly irritating." This comment related specifically to the music, and as discussed in Sec. 3.3, it is therefore important to ensure that specific metadata are included to enhance qualities such as the spatial attributes of content in a genre-specific way.

The low-level perception category also contained a number of codes that fell into their own concept. The clear sounds concept was found to have a positive effect ($N = 3$), described with comments such as object sounds being "clear and distinct." This suggests a benefit of creating distinct localizable sound events. One negative comment, however, suggested the sources were "too separated." Concepts receiving a negative response were cohesion ($N = 2$), loudness balance ($N = 2$), and timbre ($N = 1$). Negative comments for cohesion, which was mentioned alongside timbre, related specifically to the effect on music, again highlighting the need for more sophisticated metadata and rendering rules. The loudness balance responses likely relate primarily to calibration issues: in some cases "the smaller speaker sounds were lost," while at other times sources "were a bit too loud." Research is required to understand how best to overcome the practical issues around calibrating an MDO system (discussed further in Sec. 3.2), but equally on how to provide an enhanced listener experience across the listening area when, for example, proximity of loudspeakers could be an issue depending on listener position.

As well as low-level perceptual factors, a high-level perception category was identified; this had a positive overall response. The category was dominated by the concept cognition and evaluation, which comprised several codes. Codes eliciting positive responses within this concept included effect on story ($N = 4$), engagment ($N = 3$), surprising ($N = 2$), and hedonic judgments ($N = 1$), with no negative responses in each case. The positive effect on story included comments relating to how MDO "worked very well for storytelling." The engagement code related to comments of feeling involved and "part of the experience," and was grouped alongside both the effect on story and immersive/enveloping. This suggests MDO is particularly suited to producing immersive and engaging storytelling content. The surprising code related to the experience being better than expected, while the sole positive hedonic judgments comment stated that the demonstration "sounded so good," athough several general comments in the hedonic judgments response data expressed similar thoughts (e.g., "a very impressive experience" and "really cool concept").

Despite the overall positive influence on high level perception, there were also a number of negatives. In the cognition and evaluation concept these were found for the codes expectation ($N = 2$), referring to something sounding unusual or unexpected (specifically in the music program); distracting ($N = 2$), referring to the positioning of dialogue; and gimmicky ($N = 2$). A single negative comment was also attributed to the code/concept listening mode, relating again to the music content and stating that MDO "wouldn't quite work for background music, but for dedicated listening would be good." Research is required to determine the optimum rule set for creating engaging and immersive content without elements that detract from the quality of listener experience.

To fully understand the impact of MDO on perception and how changes in perceptual attributes contribute to the overall quality of listening experience, further controlled evaluation is required. It would be beneficial to compare MDO against other realistic home spatial audio systems, both quantitatively (i.e., with ratings of quality of experience or other similar attributes) and qualitatively (determining the positive and negative aspects of MDO that lead to particular ratings, in order that these aspects can be improved).

3.2 Technical

The analysis revealed a number of technical aspects relating to the delivery and implementation of MDO. The technical category had the largest number of negative comments (8 positive, 12 negative, and 10 general), suggesting that the undesirable aspects were largely technical in nature. Negative comments predominantly related to the codes listener position and quality/type of loudspeakers within the physical setup concept. Those relating to listener position ($N = 4$) were associated with how strongly dependent the experience was on listener location; statements described that the listening experience was "rather dependent on seating position," with one specific comment that "off center doesn't sound good." Intrinsically linked is the location of the loudspeakers, with negative comments relating to height of loudspeakers ("narrator voice seemed to distract from the screen as it was low between the stereo pair") and proximity of loudspeakers ("found the closer speakers annoying") respectively.

MDO could, therefore, be developed by introducing knowledge of the position of the listener(s) relative to the loudspeakers. This knowledge could be collected using a listener tracking system [42], and would require a more advanced metadata adaptation rule set. Such optimization could provide a benefit over traditional rendering methods in terms of removing the "sweet spot." It would also be beneficial to find out how listeners interact with an MDO

system; for example, if a listener is unhappy with the location of a wireless device they might simply choose to move it or adjust its volume control to produce a setup that suits their preferences. Equally, object-based audio and MDO offers opportunities for personalized content; for example, providing level-boosted speech, audio description, or objects important to the narrative to personal devices for the hearing- or visually-impaired [43]. Moving the narrator from a stereo mix to an auxiliary loudspeaker could improve speech intelligibility through spatial release from masking.

The remaining comments relating to negative physical setup aspects considered the quality and type of loudspeakers, albeit as part of a more balanced response of positive ($N = 3$) and negative ($N = 4$) comments. Negative traits related to noticeable differences in speaker "quality" and "frequency response," while positive comments noted "how effective this was given the small size of the additional speakers." There are occasions when a device will not be suitable to reproduce a given object sound. Consequently, it is important to understand both the required metadata and rendering methods to best select devices for different object types and audio signal features. This requires further investigation.

A smaller concept within the technical category described the rendering methods used. DOTLR produced positive comments for both object-to-loudspeaker mapping and adaptation due to the system's ability to be able to update in real time when loudspeakers were turned on or off. However, this method resulted in occasional "jumpy movement." Further research is required to understand how to optimally route objects to auxiliary loudspeakers, as well as to understand how to deal with movement. MDO benefits from object-based audio by utilizing different rendering methods as most appropriate for the objects and available loudspeakers. Development could focus on new rendering methods that make best use of loudspeakers of different types and qualities. One area of particular interest is in the rendering of reverberant or diffuse sound objects [44].

The practicality of MDO was mentioned as both a positive (due to "using everyday devices" and being "a great way to have surround sound at home") and a negative (due to "needing multiple devices"); the latter is seen as being a substantially lower barrier to enhanced spatial audio reproduction than the high channel count methods described in Sec. 1.1.

Further general comments raised technical challenges relating to "practicality of tracking speakers," reproduction in "non-treated home environment[s]," as well as possible delivery methods such as "over the Internet." Challenges relating the practical implementation of such a system (which include discovery and pairing, synchronization, localization, calibration, and metadata collection) are beyond the scope of this paper. However, there is a great deal of ongoing work in this area. For example, there are standards for connecting to and synchronizing second screen devices [45]; methods for synchronizing audio, video, and data over Wi-Fi [46]; toolboxes for creating ad hoc networks of mobile devices for musical performance [47]; and various indoor positioning systems (for people and objects) utilizing different technologies [48]. Implementation of technical solutions will be the focus of future work.

3.3 Content

Twenty-four codes (9 positive, 12 negative, and 3 neutral) were related specifically to the content. These primarily fell into effecton program type and effect on object type concepts. For the effect on program type the responses related to the effect on drama were all positive ($N = 7$), while the responses related to the effect on music were all negative ($N = 7$). The responses coded as effect on drama overlapped with the immersive/enveloping, effect on atmos, and clear sounds codes, suggesting that MDO is particularly suitable for immersive drama. Conversely, responses in the effect on music code overlapped with negative responses from codes including spatial image, expectation, gimmicky, and cohesion. Additional comments related to an "unusual" listening experience that "didn't feel natural" and "being used to a traditional front facing listening experience," as well as questioning the suitability for the genre of the music (pop) and whether or not other types of music might be more suited. From the analysis, it is not possible to determine whether the positive or negative experiences were engendered by the specific content items (including how they were mixed and rendered) or because of their genres.

It is also necessary to investigate genre-specific production techniques and metadata for MDO. For example, more subtle or less intrusive use of augmented devices may often be appropriate. There are unanswered questions relating to whether aversion to reproduction where there is a real-world reference (e.g., musicians performing on a stage) are inherent or due to the initially unfamiliar experience. On a practical level, there remain many questions pertaining to how a producer would go about creating content for a system with an unknown array of devices in a range of potentially variable positions. Producers may wish to attach metadata to define limits of how conservative the final rendering should be; for example, it may be useful to specify that all dialogue should remain in the front main speakers. Ethnographic studies of object-based content creation have been used as a way to find out about the experience of producers and listeners in new spatial audio systems [40]. A similar approach could be taken to developing new MDO content and learning about the production process in order to generate an optimal metadata adaptation rule set for content across a range of genres.

As well as responses relating to the full demonstrated scenes, individual aspects of the audio objects within these scenes were identified in the response data. Within the effect on object type concept, the codes effect on FX, effect on voice, and effect on atmos were identified as having a broadly positive ($N = 1$), negative ($N = 2$), and neutral ($N = 1$ positive, $N = 1$ negative) effect on the MDO experience respectively. Through use of advanced metadata, along with a rendering rule set as discussed in Sec. 1.3, it is possible to use information describing object types to determine how the scene should be reproduced and how

objects should be routed to the individual devices. Research is required into the appropriate use of semantic metadata to classify object type such that a renderer can more intelligently route differing objects in an MDO system. For example, augmented devices may be suitable for rendering atmospheric sound, which was rated positively (e.g., "loved the atmos in the drama"); however, greater understanding is required on how best to treat dialogue. In general, further work could focus on rules for rendering different types of object (for example, considering the object categories determined by Woodcock et al. [49]).

There was also one comment relating to audio-visual interaction, reporting that the "narrator voice seemed to distract from the screen." While for the demonstration the screen displayed a user interface only, the comment relates more broadly to the effect on voice code and diegetic or nondiegetic sounds. The narrator in the drama scene, for example, was routed to the auxiliary loudspeaker positioned close to and in front of the listener, and hence was spatially separated from the screen. There is an expectation that dialogue will appear from the front and/or screen direction; this is particularly true for diegetic sounds, but also for nondiegetic narration or dialogue. Informal comments following demonstrations of MDO have suggested that the narrator position splits opinion; some participants have commented that narration being replayed through an auxiliary device has a strong positive effect.

In future work, MDO could be utilized to create a multimodal experience—for example, using different visual content reproduced on devices (as in second-screen experiences [50]) as well as connected lighting or temperature systems in smart homes. Effects such as audio-visual interaction [51] therefore raise additional possibilities and challenges that require further investigation.

4 SUMMARY

A system that augmented a stereo pair of loudspeakers with an ad hoc array of connected devices was described. The MDO approach aims to optimize aspects of the listening experience that are closely related to listener preference rather than attempting to recreate sound fields as devised during production. This MDO approach cannot be expected to preserve attributes such as localization accuracy and timbral homogeneity, which have often been seen as primary factors in the quality of spatial audio systems. However, it does provide a realistic way of using loudspeakers at different positions and distances, giving the potential to increase perception of important attributes such as listener envelopment.

An MDO system was implemented using an adaptive object-based audio framework. The system relied on detailed metadata for describing the loudspeakers and audio objects, and a rule set for automatically adapting the reproduction. The system was demonstrated to 20 participants and a free text elicitation exercise was conducted. Thematic analysis was performed on the elicited text data to determine concepts that were positively or negatively related to the experience of listeners. It was shown that listeners had a positive experience due to the increased immersion compared to stereo reproduction, and that the MDO approach worked particularly well for drama content. Negative concepts were recorded for other content (music and radio advert), the different types and qualities of loudspeakers, and variations caused by listener position. However, the overall comments suggested that the listeners' experience of MDO was positive. The analysis was used to motivate suggestions for future work, particularly highlighting the need for development of the production process and metadata models, technical solutions to delivering content and establishing an ad hoc loudspeaker system, and evaluation of the listening experience.

5 ACKNOWLEDGMENTS

This work was supported by the EPSRC Programme Grant S3A: Future Spatial Audio for an Immersive Listener Experience at Home (EP/L000539/1). The authors would like to acknowledge the input of the following researchers who contributed through discussions of ideas around media device orchestration and/or commenting on drafts of the paper: Bruno Fazenda, Dylan Menzies, Philip Coleman, Hansung Kim, Qingju Liu, Marcos Simon Galvez, Hanne Stenzel, Ben Shirley, and Craig Cieciura. Additionally, some of the ideas described above were explored during a "hack week" in January 2016; the authors would like to acknowledge the participation of Frank Melchior, Yan Tang, Catherine Robinson, Sam Fowler, Miguel Blanco Galindo, and Ben Hammond. The radio advert program item was produced by Tim McKeever. One of the loudspeakers used in the demonstration system was kindly lent to the S3A project by Bang & Olufsen. Details about the data underlying this work, along with the terms for data access, are available from https://doi.org/10.17866/rd.salford.5589856.

6 REFERENCES

[1] J. Francombe, R. Mason, P. Jackson, T. Brookes, R. Hughes, J. Woodcock, A. Franck, F. Melchior, and C. Pike, "Media Device Orchestration for Immersive Spatial Audio Reproduction," *Audio Mostly Conference on Augmented and Participatory Sound and Music Experiences Proceedings* (ACM, 2017 Aug.), https://doi.org/10.1145/3123514.3123563.

[2] ITU-R rec. BS.2051, "Advanced Sound System for Programme Production," *ITU-R Broadcasing Service (Sound) Series* (2014).

[3] K. Hamasaki and K. Hiyama, "Reproducing Spatial Impression with Multichannel Audio," presented at the *AES 24th International Conference on Multichannel Audio* (2003 Jun.), conference paper 19.

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

[5] W. Howie, R. King, D. Martin, and F. Grond, "Subjective Evaluation of Orchestral Music Recording Techniques for Three-Dimensional Audio," presented at the

142nd Convention of the Audio Engineering Society (2017 May), convention paper 9797.

[6] V. Pulkki, "Virtual Sound Source Positioning Using Vector Base Amplitude Panning," *J. Audio Eng. Soc.*, vol. 45, pp. 456–466 (1997 Jun.).

[7] M. A. Gerzon, "Periphony: With-Height Sound Reproduction," *J. Audio Eng. Soc.*, vol. 21, pp. 2–10 (1973 Feb.).

[8] 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.*, vol. 62, pp. 821–830 (2015 Jan./Feb.), https://doi.org/10.17743/jaes.2014.0049.

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

[10] G. Theile and G. Plenge, "Localization of Lateral Phantom Sources," *J. Audio Eng. Soc.*, vol. 25, pp. 196–200 (1977 Apr.).

[11] R. Conetta, T. Brookes, F. Rumsey, S. Zielinski, M. Dewhirst, P. Jackson, S. Bech, D. Meares, and S. George, "Spatial Audio Quality Perception (Part 1): Impact of Commonly Encountered Processes," *J. Audio Eng. Soc.*, vol. 62, pp. 831–846 (2015 Jan./Feb.), https://doi.org/10.17743/jaes.2014.0048.

[12] ITU-R rec. BS.775-3, "Multichannel Stereophonic Sound System With and Without Accompanying Picture," *ITU-R Broadcasing Service (Sound) Series* (2012).

[13] 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.), https://doi.org/10.17743/jaes.2015.0062.

[14] H. Wierstorf, D. Ward, R. Mason, E. M. Grais, C. Hummersone, and M. D. Plumbley, "Perceptual Evaluation of Source Separation for Remixing Music," presented at the *143rd Convention of the Audio Engineering Society* (2017 Oct.), convention paper 9880.

[15] A. J. Heller and E. M. Benjamin, "The Ambisonic Decoder Toolbox: Extensions for Partial-coverage Loudspeaker Arrays," *Linux Audio Conference, Karlsruhe, Germany* (2014 May).

[16] F. Zotter, "Sampling Strategies for Acoustic Holography/Holophony on the Sphere," *Fortschritte der Akustik, NAG/35. DAGA International Conference, Rotterdam, Italy* (2009 Mar.).

[17] O. Kohsaka, E. Satoh, and T. Nakayama, "Sound-Image Localization in Multichannel Matrix Reproduction," *J. Audio Eng. Soc.*, vol. 20, pp. 542–548 (1972 Sep.).

[18] M. A. Gerzon, "Criteria for Evaluating Surround-sound Systems," *J. Audio Eng. Soc.*, vol. 25, pp. 400–408 (1977 Jun.).

[19] G. Martin, W. Woszczyk, J. Corey, and R. Quesnel, "Sound Source Localization in a Five-channel Surround Sound Reproduction System," presented at the *107th Convention of the Audio Engineering Society* (1999 Sep.), convention paper 4994.

[20] V. Pulkki, M. Karjalainen, and V. Välimäki, "Localization, Coloration, and Enhancement of Amplitude-panned Virtual Sources," presented at the *AES 16th lnternational Conference on Spatial Sound Reproduction* (1999 Mar.), conference paper 16-024.

[21] M. Poletti, "Robust Two-Dimensional Surround Sound Reproduction for Nonuniform Loudspeaker Layouts," *J. Audio Eng. Soc.*, vol. 55, pp. 598–610 (2007 Jul./Aug.).

[22] L. S. Simon, R. Mason, and F. Rumsey, "Localization Curves for a Regularly-Spaced Octagon Loudspeaker Array," presented at the *127th Convention of the Audio Engineering Society* (2009 Oct.), convention paper 8079.

[23] R. Wallis and H. Lee, "The Effect of Interchannel Time Difference on Localization in Vertical Stereophony," *J. Audio Eng. Soc.*, vol. 63, pp. 767–776 (2015 Oct.), https://doi.org/10.17743/jaes.2015.0069.

[24] S. Bertet, J. Daniel, E. Parizet, and O. Warusfel, "Investigation on Localisation Accuracy for First and Higher Order Ambisonics Reproduced Sound Sources," *Acta Acustica United With Acustica*, vol. 99, pp. 642–657 (2013 Jul./Aug.), https://doi.org/10.3813/AAA.918643.

[25] 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.), https://doi.org/10.17743/jaes.2016.0071.

[26] 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. Am.*, vol. 117, pp. 3832–3840 (2005 Jun.), https://doi.org/10.1121/1.1904305.

[27] R. Mason, "How Important Is Accurate Localization in Reproduced Sound?" presented at the *142nd Convention of the Audio Engineering Society* (2017 May), convention paper 9759.

[28] J. Berg, "The Contrasting and Conflicting Definitions of Envelopment," presented at the *126th Convention of the Audio Engineering Society* (2009), convention paper 7808.

[29] S. George, F. Rumsey, S. Zielinski, and S. Bech, "Evaluating the Sensation of Envelopment Arising from 5-Channel Surround Sound Recordings," presented at the *124th Convention of the Audio Engineering Society* (2008 May), convention paper 7382.

[30] 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.), https://doi.org/10.17743/jaes.2016.0070.

[31] T. Walton, M. Evans, D. Kirk, and F. Melchior, "A Subjective Comparison of Discrete Surround Sound and Soundbar Technology by Using Mixed Methods," presented at the *140th Convention of the Audio Engineering Society* (2016 Jun.), convention paper 9592.

[32] J.-M. Jot, V. Larcher, and O. Warusfel, "Digital Signal Processing Issues in the Context of Binaural and Transaural Stereophony," presented at the *98th Convention of the Audio Engineering Society* (1995 Feb.), convention paper 3980.

[33] MPEG, "Context and Objectives for Media Orchestration v.3," Tech. rep., The Moving Picture Ex-

perts Group W16131 (https://mpeg.chiariglione.org/sites/default/files/files/standards/parts/docs/W16131 (2016 Feb.).

[34] S. Füg, A. Hoelzer, C. Borss, C. Ertel, M. Kratschmer, and J. Plogsties, "Design, Coding and Processing of Metadata for Object-based Interactive Audio," presented at the *137th Convention of the Audio Engineering Society* (2014 Oct.), convention paper 9097.

[35] H. Purnhagen, T. Hirvonen, L. Villemoes, J. Samuelsson, and J. Klejsa, "Immersive Audio Delivery Using Joint Object Coding," presented at the *140th Convention of the Audio Engineering Society* (2016 May), convention paper 9587.

[36] P. Coleman, A. Franck, J. Francombe, Q. Liu, T. de Campos, R. Hughes, D. Menzies, M. Simon Galvez, Y. Tang, J. Woodcock, P. Jackson, F. Melchior, C. Pike, F. Fazi, T. Cox, and A. Hilton, "An Audio-Visual System for Object-Based Audio: From Recording to Listening," IEEE Transactions on Multimedia (2018), https://doi.org/10.1109/TMM.2018.2794780.

[37] ITU-R rec. BS.2088, "Long-Form File Format for the International Exchange of Audio Programme Materials with Metadata," *ITU-R Broadcasing Service (Sound) Series* (2015).

[38] ITU-R rec. BS.2388-1, "Usage Guidelines for the Audio Definition Model and Multichannel Audio Files," *ITU-R Broadcasing Service (Sound) Series* (2017).

[39] ITU-R rec. BS.1116-3, "Methods for the Subjective Assessment of Small Impairments in Audio Systems," *ITU-R Broadcasing Service (Sound) Series* (2015).

[40] J. Woodcock, C. Pike, F. Melchior, P. Coleman, A. Franck, and A. Hilton, "Presenting the S3A Object-Based Audio Drama Dataset," presented at the *140th Convention of the Audio Engineering Society,* (2016 Jun.), e-Brief 255.

[41] V. Braun, and V. Clarke, "Using Thematic Analysis in Psychology," *Qualitative Research in Psychology*, vol. 3, pp. 77–101 (2006 Jul.), https://doi.org/10.1191/1478088706qp063oa.

[42] Q. Liu, T. de Campos, W. Wang, P. Jackson, and A. Hilton, "Person Tracking Using Audio and Depth Cues," *IEEE International Conference on Computer Vision (ICCV) Workshops* (2015 Dec.), https://doi.org/10.1109/ICCVW.2015.97.

[43] B. G. Shirley, M. Meadows, F. Malak, J. S. Woodcock, and A. Tidball, "Personalized Object-based Audio for Hearing Impaired TV Viewers," *J. Audio Eng. Soc.*, vol. 65, pp. 293–303 (2017 Apr.), https://doi.org/10.17743/jaes.2017.0005.

[44] P. Coleman, A. Franck, P. J. Jackson, R. J. Hughes, L. Remaggi, and F. Melchior, "Object-Based Reverberation for Spatial Audio," *J. Audio Eng. Soc.*, vol. 65, pp. 66–77 (2017 Jan./Feb.), https://doi.org/10.17743/jaes.2016.0059.

[45] DVB, "Companion Screens and Streams—Expanding the TV Experience with Companion Screens," Tech. rep., DVB Fact Sheet (https://www.dvb.org/resources/public/factsheets/dvb-css_factsheet.pdf) (2017 Jul.).

[46] Wi-Fi Alliance, "Wi-Fi TimeSync," https://www.wi-fi.org/discover-wi-fi/wi-fi-timesync (accessed: 17 Nov. 2017) (2017).

[47] S. Robaszkiewicz and N. Schnell, "Soundworks—A Playground for Artists and Developers to Create Collaborative Mobile Web Performances," *1st Web Audio Conference, Paris, France* (2015 Jan.).

[48] A. Correa, M. Barcelo, A. Morell, and J. L. Vicario, "A Review of Pedestrian Indoor Positioning Systems for Mass Market Applications," *Sensors*, vol. 17, pp. 1927–1953 (2017 Aug.), https://doi.org/10.3390/s17081927.

[49] J. Woodcock, W. J. Davies, T. J. Cox, and F. Melchior, "Categorization of Broadcast Audio Objects in Complex Auditory Scenes," *J. Audio Eng. Soc.*, vol. 64, pp. 380–394 (2016 Jun.), https://doi.org/10.17743/jaes.2016.0007.

[50] M. Bober, I. Feldmann, S. G. Lobo, A. Messina, S. Paschalakis, G. Perrone, V. Scurtu, and G. Vavalà, "BRIDGET: An Approach at Sustainable and Efficient Production of Second Screen Media Applications," *IBC 2015: International Broadcasting Convention, Amsterdam, Netherlands* (2015 Sep.), https://doi.org/10.1049/ibc.2015.0001.

[51] A. Kohlrausch and S. van de Par, "Audio-Visual Interaction in the Context of Multi-Media Applications," J. Blauert (ed.), *Communication Acoustics*, pp. 109–138 (Springer, 2005), https://doi.org/10.1007/b139075.

THE AUTHORS


Jon Francombe


James Woodcock


Richard Hughes


Russell Mason


Andreas Franck


Chris Pike


Tim Brookes


Bill Davies


Philip Jackson


Trevor Cox


Filippo Maria Fazi


Adrian Hilton

Jon Francombe graduated with a first-class honours degree in music and sound recording (Tonmeister) from the University of Surrey, Guildford, UK, in 2010, and received a Ph.D. in perceptual audio quality evaluation from the same institution in 2014. He then worked as a research fellow on the EPSRC-funded "S3A: Future Spatial Audio for an Immersive Listener Experience at Home" project, investigating the perceptual attributes of spatial audio reproduction and new methods for immersive audio reproduction. Jon currently works as a senior research and development engineer in the audio team at BBC R&D.

•

James Woodcock is a research fellow at the University of Salford. His primary area of research is the perception and cognition of complex sound and vibration. James holds a B.Sc. in audio technology, an 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. James is currently working on the S3A project. His work mainly focuses on the perception of auditory objects in complex scenes, the listener experience of spatial audio, and intelligent rendering for object-based audio.

•

Richard Hughes received a B.Sc. first-class honours degree in audio technology from the University of Salford, UK, in 2006, before completing a Ph.D. at the same institution in 2011 in the area of architectural acoustics. From 2012 to 2014 he worked as a Research Associate at the University of Manchester, UK. He is currently working in the Acoustics Research Centre at Salford as a research fellow as part of the S3A project. His research interests include among others: room acoustics; acoustic modeling of rooms, diffuser design and application; array theory and multichannel systems; binaural and auralization; rendering and perception of spatial audio reproduction; and digital signal processing.

•

Russell Mason graduated from the University of Surrey in 1998 with a B.Mus. in music and sound recording (Tonmeister). He was awarded a Ph.D. in audio engineering and psychoacoustics from the University of Surrey in 2002, and was subsequently employed as a research fellow. He is currently a senior lecturer in the Institute of Sound Recording, University of Surrey, and is program director of the undergraduate Tonmeister program. Russell's research interests are focused on psychoacoustic engineering, including the development of methods for subjective evaluation, and modeling aspects of auditory perception.

•

Andreas Franck received the Diploma degree in computer science and the Ph.D. degree in electrical engineering, both from the Ilmenau University of Technology, Germany. Since 2004 he has been with the Fraunhofer Institute for Digital Media Technology IDMT, Ilmenau, Germany. In 2014 he joined the Institute of Sound and Vibration research, University of Southampton, UK as a postdoctoral research fellow. He is currently working in the S3A project. His research interests include spatial and object-based audio, efficient reproduction algorithms, audio

signal processing, and architecture and implementation of audio software. Dr. Franck is a member of IEEE, IEEE Signal Processing Society, and the Audio Engineering Society.

•

Chris Pike leads the audio research team at BBC R&D, which focuses on the technology for immersive and personalized listening experiences. He leads the BBC Audio Research Partnership, through which the BBC collaborates with world-class universities on audio research, and is active in industry bodies such as the ITU and the EBU. His work has seen the BBC using spatial audio with some of its biggest program brands, such as the "Proms," "Planet Earth," and "Doctor Who." He is also a Ph.D. candidate in the Audio Lab at the University of York, investigating the quality of binaural audio systems.

•

Tim Brookes received the B.Sc. degree in mathematics and the M.Sc. and D.Phil. degrees in music technology from the University of York, York, UK, in 1990, 1992, and 1997, respectively. He was employed as a software engineer, recording engineer, and research associate before joining, in 1997, the academic staff at the Institute of Sound Recording, University of Surrey, Guildford, UK, where he is now Senior Lecturer in Audio and Director of Research. His teaching focuses on acoustics and psychoacoustics and his research is in psychoacoustic engineering: measuring, modeling, and exploiting the relationships between the physical characteristics of sound and its perception by human listeners.

•

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.

•

Philip Jackson is senior lecturer in machine audition at the Centre for Vision, Speech & Signal Processing (CVSSP, University of Surrey, UK) with MA in engineering (Cambridge University, UK), and Ph.D. in electronic engineering (University of Southampton, UK). His broad interests in acoustical signal processing have led to research contributions in speech production, auditory processing and recognition, audio-visual machine learning, blind source separation, articulatory modeling, visual speech synthesis, spatial audio sound recording, reproduction and quality evaluation, and sound field control [Google Scholar: bit.ly/2oTRw1C]. He leads research on object-based production in the S3A project and enjoys many kinds of listening experience.

•

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 award 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 24 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* will be published in May 2018. @trevor cox.

•

Filippo Maria Fazi graduated in mechanical engineering from the University of Brescia (Italy) in 2005. He obtained his Ph.D. in acoustics from the Institute of Sound and Vibration Research (ISVR) of the University of Southampton, UK, in 2010, with a thesis on sound field reproduction. In the same year he was awarded a fellowship by the Royal Academy of Engineering and by the Engineering and Physical Sciences Research Council. He is currently an associate professor at the University of Southampton. Dr. Fazi's research interests include audio technologies, electroacoustics, and digital Signal Processing, with special focus on acoustical inverse problems, multichannel systems, virtual acoustics, microphone, and loudspeaker arrays.

•

Adrian Hilton is Director of the Centre for Vision, Speech and Signal Processing at the University of Surrey. Since joining CVSSP in 1992, he has published more than 200 papers in leading journals and conferences. He has received two EU IST Prizes, a DTI Manufacturing Industry Achievement Award, and a Computer Graphics World Innovation award; eight best paper awards in international journals and conferences; and co-founded the European Conference on Visual Media Production. He holds a Royal Society Wolfson Research Merit Award (2013–18) and held a Royal Society Industry Fellowship (2008–11) with Framestore to investigate multi-view and 3D video in film production. Prof. Hilton is principal investigator on the S3A project.