OPEN ACCESS Freely available online
(cc) BY

J. Francombe, T. Brookes, and R. Mason, "Determination and Validation of Mix Parameters for Modifying Envelopment in Object-Based Audio"
J. Audio Eng. Soc., vol. 66, no. 3, pp. 127–145, (2018 March.).
DOI: https://doi.org/10.17743/jaes.2018.0011

Determination and Validation of Mix Parameters for Modifying Envelopment in Object-Based Audio

JON FRANCOMBE,* *AES Associate Member*, **TIM BROOKES,** *AES Member*, **AND**

RUSSELL MASON, *AES Member*

Institute of Sound Recording, University of Surrey, Guildford, UK

Envelopment is an important attribute of listener preference for spatial audio reproduction. Object-based audio offers the possibility of altering the rendering of an audio scene in order to modify or maintain perceptual attributes—including envelopment—if the relationships between attributes and mix parameters are known. In a method of adjustment experiment, mixing engineers were asked to produce mixes of four program items at low, medium, and high levels of envelopment in 2-channel, 5-channel, and 22-channel reproduction systems. The participants could vary a range of level, position, and equalization parameters that can be modified in object-based audio systems. The parameters could be varied separately for different semantic object categories. Nine parameters were found to have significant relationships with envelopment; parameters relating to the horizontal and vertical spread of sources were shown to be most important. A follow-on experiment demonstrated that these parameters can be adjusted to produce a range of envelopment levels in other program items.

0 INTRODUCTION

Object-based audio is widely seen as an important format for current and future spatial audio creation and distribution [1]; various systems have been proposed or described in the literature [2–5]. In object-based audio transmission, a scene is distributed as a set of audio objects, as opposed to the loudspeaker feeds that are distributed in channel-based audio. An object comprises an audio stream for a particular aspect of the scene, accompanied by some metadata (such as the desired level and spatial position of the object). For example, a string quartet scene might contain four audio objects—violin one, violin two, viola, and cello. Loudspeaker feeds (for any arbitrary loudspeaker layout) are generated by the *renderer* based on the audio streams and metadata (normally using some panning algorithm, e.g., vector-base amplitude panning (VBAP) [6]). This provides various opportunities for modifying the creation of the loudspeaker feeds to provide an optimal listening experience. For example, the number of loudspeakers and their positions can be accounted for, and listener interaction can be facilitated [7, 8].

This disconnect between content and loudspeaker feed generation also provides possibilities for perceptual optimization. It is conceivable that changes can be made in metadata or rendering algorithms to change the listening experience in a perceptually optimized manner. This relies on knowledge of the perceptual attributes that should be considered and of the relationships between those attributes and relevant metadata parameters so that appropriate changes can be made in the rendering.

There is a large body of research concerning the important perceptual attributes of spatial audio reproduction; see, for example, Francombe et al. [9] for a review. Francombe et al. [10] found that "envelopment" was the most important attribute of listener preference for spatial audio reproduction systems. This supports findings made in previous literature about the importance of envelopment in reproduced audio (e.g., [11]) as well as the well-known importance of listener envelopment in concert hall acoustics [12]. It would therefore be beneficial to develop a tool that could modify envelopment in object-based audio in a perceptually informed manner. Such a tool would form an extension of the personalization controls provided in existing object-based audio systems, and/or be a means by which a renderer could maintain a desired level of envelopment when feeding a sub-optimal loudspeaker array.

A number of existing models of envelopment in reproduced sound were reviewed (Sec. 1.2). None of the existing

*Now at BBC Research and Development, MediaCityUK, Salford, UK. To whom correspondence should be addressed: jon.francombe@bbc.co.uk

models were found to be suitable for use in adapting object-based audio reproduction to control envelopment. Consequently, an investigation of envelopment in object-based audio was conducted in order to determine perceptual rules that could be used to produce a tool that can modify envelopment in object-based audio.

Following the literature review in Sec. 1, the design of an experiment to determine the relationship between parameters of an object-based music mix and perceived envelopment is presented (Sec. 2). In Sec. 3 the results are analyzed in order to determine the relationship between the parameters tested and the level of envelopment produced. A validation of the results is presented in Sec. 4, in which it is shown that the parameters determined can be used to modify envelopment in new stimuli. In Sec. 5 the findings are discussed and suggestions for further work are presented.

1 ENVELOPMENT IN THE LITERATURE

Due to the perceptual importance of envelopment there has been a large amount of work on defining the attribute, identifying the relevant physical parameters, and developing predictive models. In Sec. 1.1, a discussion of the definition of envelopment is presented. In Sec. 1.2 previous attempts to model envelopment are reviewed. The aims of the current study are outlined in Sec. 1.3.

1.1 Definition of Envelopment

As noted above, envelopment is widely held to be an important factor of listener experience. However, as with any such high-level attribute, it is difficult to agree on a universal definition—particularly when a term may mean different things to different communities. This makes it important to clearly define the term when it is used in any perceptual tests.

The definition of envelopment has been considered by a number of authors [13–15]; consequently, an exhaustive literature review to find all of the definitions that have been used is unnecessary here. Berg [13] notes that there is a set of attributes with similar or overlapping meanings in the literature (listing terms such as spaciousness, spatial impression, listener envelopment (LEV), immersion/immersiveness/spatial immersion, sense of space, and surroundness). However, he also states that despite the apparent differences in definitions, there are often similarities between experiment results, suggesting that there is an underlying percept that listeners understand and agree on. Other authors have also considered the overlap between terms; for example, Griesinger [16] equates envelopment with spaciousness or spatial impression.

There is seemingly a difference between envelopment in concert halls and in reproduced audio. In concert halls, the term LEV is often used; LEV is mainly related to late lateral reflections [14]. In the context of room acoustics, LEV is related to—and often discussed alongside—apparent source width. Apparent source width (ASW) is defined as *"a broadening of the apparent width of the sound source"* [16, 17] and related to lateral energy fraction and interaural cross correlation (IACC) [18]. Rumsey [15] states that the difference between natural and reproduced sound is great enough that the important perceptual attributes differ between the two listening modes; the research presented in this paper focuses on reproduced audio.

In reproduced audio, envelopment can be produced either by multiple direct sounds being reproduced from different angles around the listener or by ambient, reverberant, or decorrelated sounds [14]. Conetta et al. [19] offers definitions of direct envelopment (*"the sensation of being surrounded by dry sources"*) and indirect envelopment (*"the sensation of being surrounded by reverberant energy or acoustic reflections"*). George et al. [14] note that the concept of direct envelopment exists in natural listening as well as in reproduced sound; for example, the sound of rain or a crowd.

In this study the definition elicited by Francombe et al. [20] was used: "*how immersed/enveloped you feel in the sound field* (from *not at all enveloping* to *fully enveloping*)." This definition is not particularly detailed—for instance, it does not differentiate between direct and indirect envelopment; however, it was produced by a panel of experienced listeners and found to relate strongly to listener preference in subsequent analysis [10]. It is similar to Berg's [13] encompassing definition of "surroundedness" (a multidimensional attribute describing *"the notion of being surrounded by sound, regardless of the sound characteristics"*).

1.2 Envelopment Models in the Literature

As envelopment is an important parameter in the listening experience of reproduced and live audio, a number of attempts have been made to develop predictive models by relating physical parameters of a reproduced sound field to subjective envelopment scores. Various models for quantifying LEV and ASW in concert halls (see van Dorp Schuitman et al. [21] for an extensive review) or virtual acoustic environments [22] have been developed. Such models either calculate parameters from a binaural room impulse response, or estimate the direct and reverberant components of a sound from the output of a binaural model. Hence, they cannot be directly applied to reproduced sound and are consequently not reviewed here. The models that aim to predict perceived envelopment for reproduced audio were reviewed and were, broadly speaking, found to fall into two categories: single metrics and complex models.

Soulodre et al. [17] developed the LG_{perc} and GS_{perc} metrics to quantify *"the level and angular distribution of the late arriving sound"*; both metrics showed a good fit to subjective training data ($r > 0.90$). Dewhirst [23] developed two metrics for envelopment prediction. The "hull metric" used the Supper localization model [24] to predict source angles and then calculated the area of the convex hull of the source angles. The "c90" metric was simply calculated as the absolute angle of the source closest to 90 degrees. The metrics showed a reasonable fit to the training data ($r = 0.79$, RMSE $= 13.73\%$ for the hull metric, and $r = 0.77$, RMSE $= 14.51\%$ for c90). Power et al. [25] investigated the envelopment produced by loudspeaker systems

with height channels, modelling the subjective results using IACC determined from binaural recordings. Again, the model showed a reasonable fit to the results with $|r| > 0.80$ for all program items.

More complex models with multiple features have also been developed. Conetta [26] fitted a model with five features calculated from various signal representations (binaural, single microphone, and microphone array signals). For different training data sets, the model performed well, with r between 0.89 and 0.96, and RMSE between 5.94 and 11.54. Dewhirst [23] also trained complex models with combinations of the metrics introduced above and similar metrics to those used by Conetta, which showed similar performance. George et al. [14] trained an envelopment model on subjective scores of envelopment collected for commercially-available 5.1 surround sound music and film excerpts, 2-channel stereo and mono, and various degradations (such as low-bit-rate coding, bandwidth limitation, and downmixing). As with the Conetta and Dewhirst models, a range of features from different signal representations was used; George et al.'s [14] model (discussed further in Sec. 4.2) used five features from loudspeaker channel feeds and a binaural representation and again showed a good fit to the training data ($r = 0.90$, RMSE $= 8.54\%$).

More complex models tend to show a better fit to their training data; however, the increased number of features often results in a reduction in generalizability to stimuli outside of those on which the models were trained.

There has been little investigation of envelopment in surround sound systems with height channels, although Power et al. [25] found that IACC showed a reasonably good correlation to perceived envelopment for systems with height channels.

The envelopment models and metrics reviewed above are suited to metering applications rather than to use in envelopment modification or perceptual optimization. For modification or optimization, the features used by the models must be directly controllable and, particularly for some of the more abstract features measured or calculated from the sound field, this might be difficult or impossible. In object-based audio, however, control of some features—those encoded in metadata—is trivial. A model of envelopment solely in terms of metadata parameters would therefore be ideal for this application. Development of such a model requires establishing the relationship between metadata parameters and the resulting envelopment.

1.3 Research Aims

Envelopment has been shown to be an important perceptual attribute of spatial audio reproduction. However, as highlighted in the literature review and discussion above, there is no current model that can easily be used to inform optimization of envelopment in reproduced audio. Consequently, an experiment was designed to investigate the parameters of an object-based mix that contribute to envelopment. The goals of the experiment were as follows.

1. To determine the relationship between parameters of an object-based mix and the perception of envelopment.
2. To develop and test a system for manipulating envelopment in object-based audio in a perceptually relevant manner.

The experiment described below also covered a range of reproduction methods; this enabled consideration of differences in the way envelopment is produced for loudspeaker systems with different characteristics (for example, those with loudspeakers behind and/or above the listener). It is also possible that individual mixing engineers produce envelopment in different ways; it was necessary to assess this before deriving generalizable rules from the data.

2 EXPERIMENT DESIGN

An experiment was designed to address the research aims outlined in Sec. 1.3. A method of adjustment task was used in which participants were asked to change parameters of a mix in order to create versions of that mix at three levels of envelopment: low, medium, and high. The following instructions were presented to the participants.

> *We are investigating the attribute of envelopment in mixes. Envelopment is defined as follows:* "how immersed/enveloped you feel in the sound field," *from* fully enveloping *to* not at all enveloping. *Your task is to produce three mixes of a program item, each with different levels of envelopment: one mix that is as enveloping as possible, one mix that has as little envelopment as possible, and one mix that is as close as possible to halfway between the two in terms of envelopment. You are also asked to keep the overall mix quality at an acceptable level in all cases.*

Participants were also instructed that they would perform this task for three reproduction systems (Sec. 2.3) and four program material items (Sec. 2.4). In total, each participant created 36 mixes: three envelopment levels × three reproduction methods × four program material items. The mixes for the three envelopment levels were created on the same test page by switching between tabs (see Fig. 1).

One test block consisted of one program material item for each of the reproduction methods. The four test blocks were performed in two sessions (i.e., two blocks—or program items—per session), which lasted about one hour each. The presentation order of the three reproduction methods within each block was randomized, and the presentation order of the blocks (i.e., the program material items) was also randomized. The random orders were different for all participants.

2.1 Test Interface and Mixing Control

The test was conducted using a hardware control surface with motorized faders. The faders, rotary controls, and buttons were mapped to parameters, and a graphical user interface (shown in Fig. 1) reported when any of the controls were moved. Mixes for the three envelopment levels

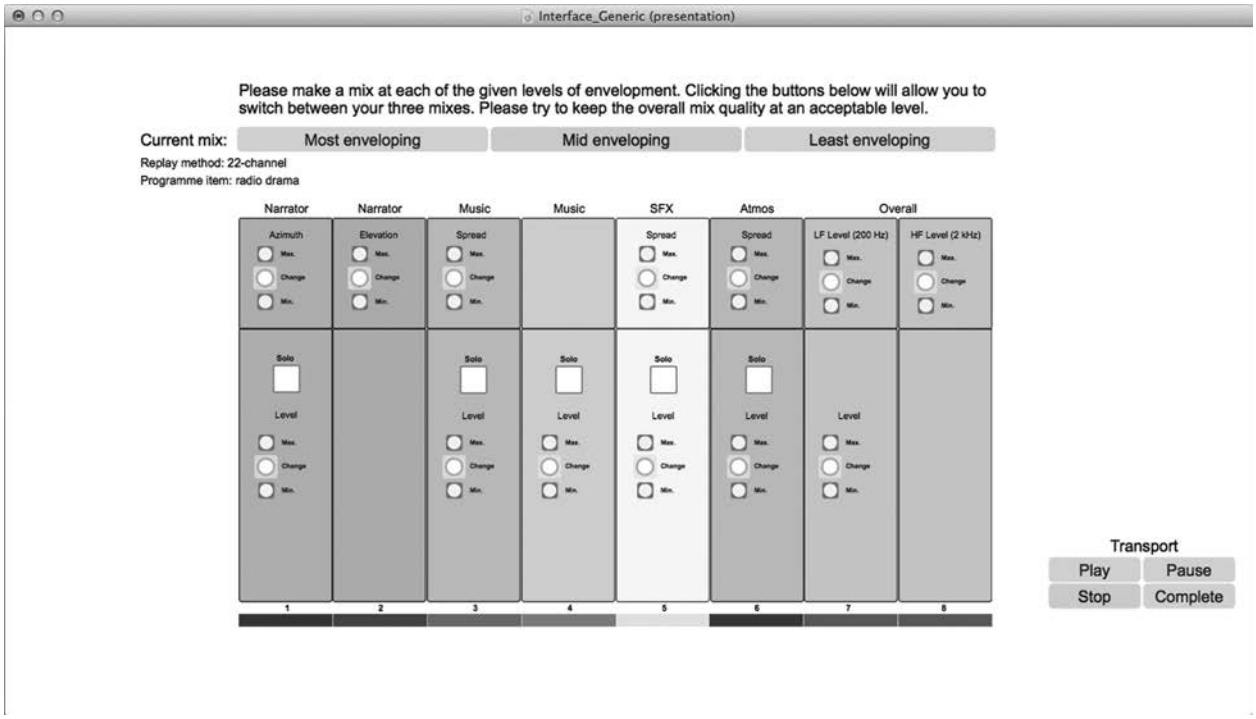


Fig. 1. Interface for the method of adjustment envelopment mixing task (radio drama program material)

were created on different tabs of the same test page; when a participant switched between tabs on the user interface (i.e., between envelopment levels), the positions of the faders on the control surface were automatically changed to match the last-stored parameters for the selected mix level.

To discourage participants from simply setting the parameter values to known positions, the controller positions were not indicated on the interface; however, to avoid confusion and frustration—especially with the continuous rotary faders—the interface indicated if any of the parameters were at their minimum or maximum values, as well as indicating when a change had been made.

The starting values of the parameters were randomized within small offsets from the default positions—again, to encourage careful listening and avoid the controller values being set to known positions. As the parameters were different for each program material item, the controller mapping was also different. In order to make the test as simple as possible for participants, the channels on the hardware control surface were color-coded to match the interface displayed on the screen. If a fader was moved but it was not associated with a parameter, it would immediately return to the bottom position.

Participants were not able to move on from a test page unless they had pressed play and spent a minimum of 20 seconds on each mix. The "complete" button was also disabled if any of the group objects was soloed; participants were informed that the solo buttons could be used to aid in making a mix but should not be used as part of the mix, and that the solo button status would not be saved in the results.

The user interface software also handled the replay of audio data and transport control. All of the parameters (with the exception of equalization) were modified by changing metadata; values were sent using Open Sound Control (OSC) messages from the interface to the *Metadapter*, a software package designed in the S3A project[1] for communicating metadata changes to the renderer. Equalization was performed on the audio objects in Max/MSP. Rendering (using VBAP) was performed using the S3A project *Versatile Interactive Scene Renderer (VISR)*. A system diagram showing the user control, audio playback, metadata adaptation, and rendering is given in Fig. 2.

2.2 Participants

The experiment was performed by participants who had at least some experience in audio mixing and did not report any significant hearing impairments. All of the participants were students or staff from the Institute of Sound Recording or the Centre for Vision, Speech, and Signal Processing at the University of Surrey.

A short questionnaire was completed by each participant prior to commencing the test. The following questions were asked.

- Please select your level of mixing experience:
  0. No previous experience
  1. Amateur/hobbyist
  2. Formal training—further education (e.g., music technology A level)
  3. Formal training—higher education (e.g., degree course)
  4. Some experience mixing in professional recording studios (e.g., internship or placement)
  5. Professional

[1] http://www.s3a-spatialaudio.org.

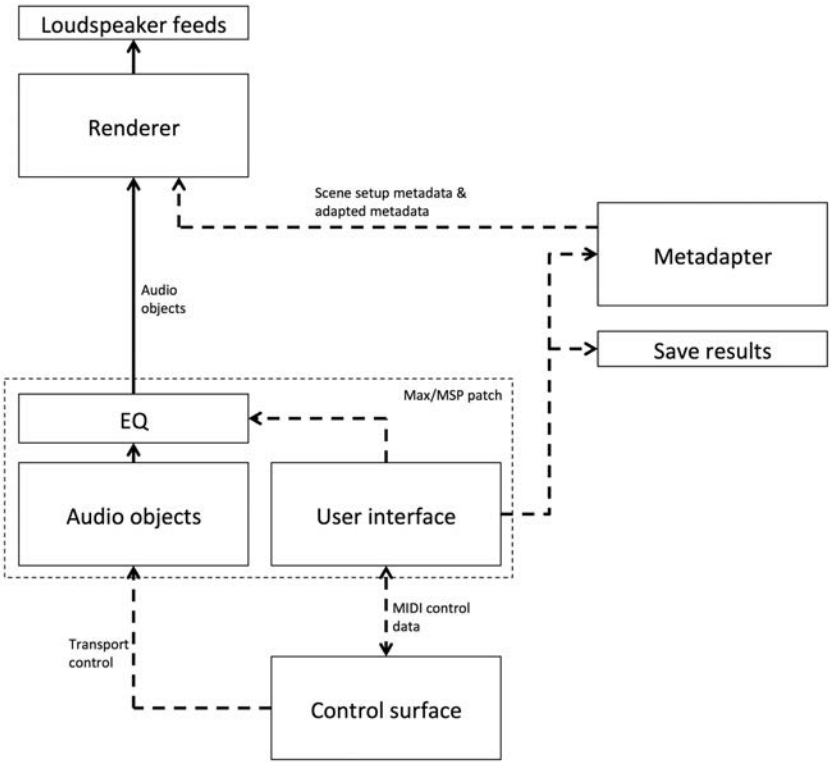


Fig. 2. System diagram for envelopment method-of-adjustment listening test

- Number of years mixing experience
- Have you made mixes for reproduction systems with more than two loudspeakers?
  - If so, please give very brief details
- Please select your level of experience of participating in listening tests:
  0. No previous experience
  1. Some experience (e.g., participation in one or two listening tests)
  2. Very experienced (e.g., participation in a wide range of listening tests in different scenarios)

The demographic data given by the participants is detailed in Table 1. A summary of the demographic data is shown in Fig. 3. All participants had at least two years mixing experience (median five years). Five participants had previous experience with surround sound mixing; all of this experience was with 5-channel systems. All participants had previously participated in listening tests.

Participants were given a £10 gift voucher as an honorarium on completion of the test.

2.3 Reproduction Systems

In order to determine how participants produced the different levels of envelopment for different reproduction systems, a variety of systems were used: 2-channel stereo, 5-channel surround sound, and 22-channel surround sound. These methods were selected as they are standardized reproduction methods [27] that are commonly used in domestic, professional, and/or research contexts and offer different degrees of potential for creating envelopment. The 2-channel stereo system can be used to create phantom images within the range ± 30 degrees; the 5-channel system offers the potential for 360 degree horizontal panning; and the 22-channel system includes loudspeakers at different heights above and below the listener.

The reproduction systems were set up in the Institute of Sound Recording listening room at the University of Surrey, an ITU-R BS.1116 standard listening room (dimensions 7.35 m $\times$ 5.70 m $\times$ 2.5 m) with a 22.2 system installed [28]. The loudspeakers used were Genelec 8330As, fed by a MOTU 24Ao audio interface. Bass management was performed using a combination of routing in the audio interface and the signal processing built into the Genelec loudspeakers; the low frequency content of each of the main channels was sent to the closer of the two subwoofers (Genelec 7350As). Aside from bass management, no subwoofer content (i.e., LFE channel) was used.

It should be noted that the reproduction system available did not have loudspeakers at the positions specified in ITU-R rec. BS.775-3 [29] for the 5-channel surround sound left surround (LS) and right surround (RS) channels. Consequently, the LS and RS loudspeakers in the 5-channel reproduction method were positioned at ± 135 degrees (rather than ± 100 to ± 120 degrees as specified in the standard).

2.4 Program Material

As in any listening test where the data will be used to make generalizations (for instance, training a predictive model), the selection of suitable program material is

Table 1. Participant demographic information. The values for mixing experience (mix. exp.) and listening test experience (test exp.) are described in the body text.

Part.	Mix. exp.	Mix. years	Surr. exp.	Details	Test exp.
1	4	2	No		2
2	3	5	Yes	A few 5.1 mixes	2
3	4	3	Yes	5.1 TV drama; 5.1 Horror short film; 5.1 TV documentaries	1
4	4	4	No		2
5	3	7	Yes	5.1 mix—mainly putting reverb & fx in the rear to open the image	2
6	3	5	No		1
7	3	5	No		1
8	4	6	Yes	Mixes for 5.1 to 22.2, object + ambisonics	2
9	1	2	No		1
10	4	5	No		1
11	4	20	No		1
12	3	4	No		2
13	2	3	No		1
14	3	2	No		2
15	4	5	No		1
16	3	4	No		2
17	1	10	Yes	Made mixes for 5.1 channel system	2

very important. It is not possible to cover every potential type of content; therefore, it is necessary to determine criteria for selection of program items. In this case, stimuli that are representative of potential object-based broadcast content were included. An experiment performed by Woodcock et al. [30] suggested that broadcast audio objects could be placed into seven categories: sounds indicating actions and movement; continuous background sounds; transient background sounds; clear speech; non-diegetic music and effects; sounds indicating the presence of people; and prominent attention-grabbing transient sounds.

Program items were selected in order to cover all of these audio object categories, as well as a range of content types and genres. The excerpts were between 20 and 35 seconds long (truncated at suitable points) with a 0.5 s linear fade in and fade out. The four program items used were as follows.

1. Radio drama: an excerpt from *The Turning Forest* scene from the S3A object-based audio drama dataset [31], featuring narration, non-diegetic music, sound effects, and background atmosphere. Objects with movement were remixed into static positions.
2. Football match: broadcast audio (captured during the FascinatE project [32]) for a football match including commentary, pitch sounds, and crowd noise. The commentator was panned centrally. An array of pitch microphones were panned to the appropriate angles. The crowd noise was captured using a Soundfield microphone and an ambisonic decode to a 10-channel subset of the 22-channel reproduction system was produced.
3. Pop track: an excerpt from an object-based remix of a pop track created for the S3A project (*Just Another Frame* by The Hotel Whisky Foxtrot [33]). As above, objects with movement were remixed into static positions.
4. Jazz duet: piano and double bass captured during the spatial audio reproduction session described by Francombe et al. [33]. *Hymn to Freedom* by Oscar Peterson and performed by Will Todd and Gareth Huw Davies.

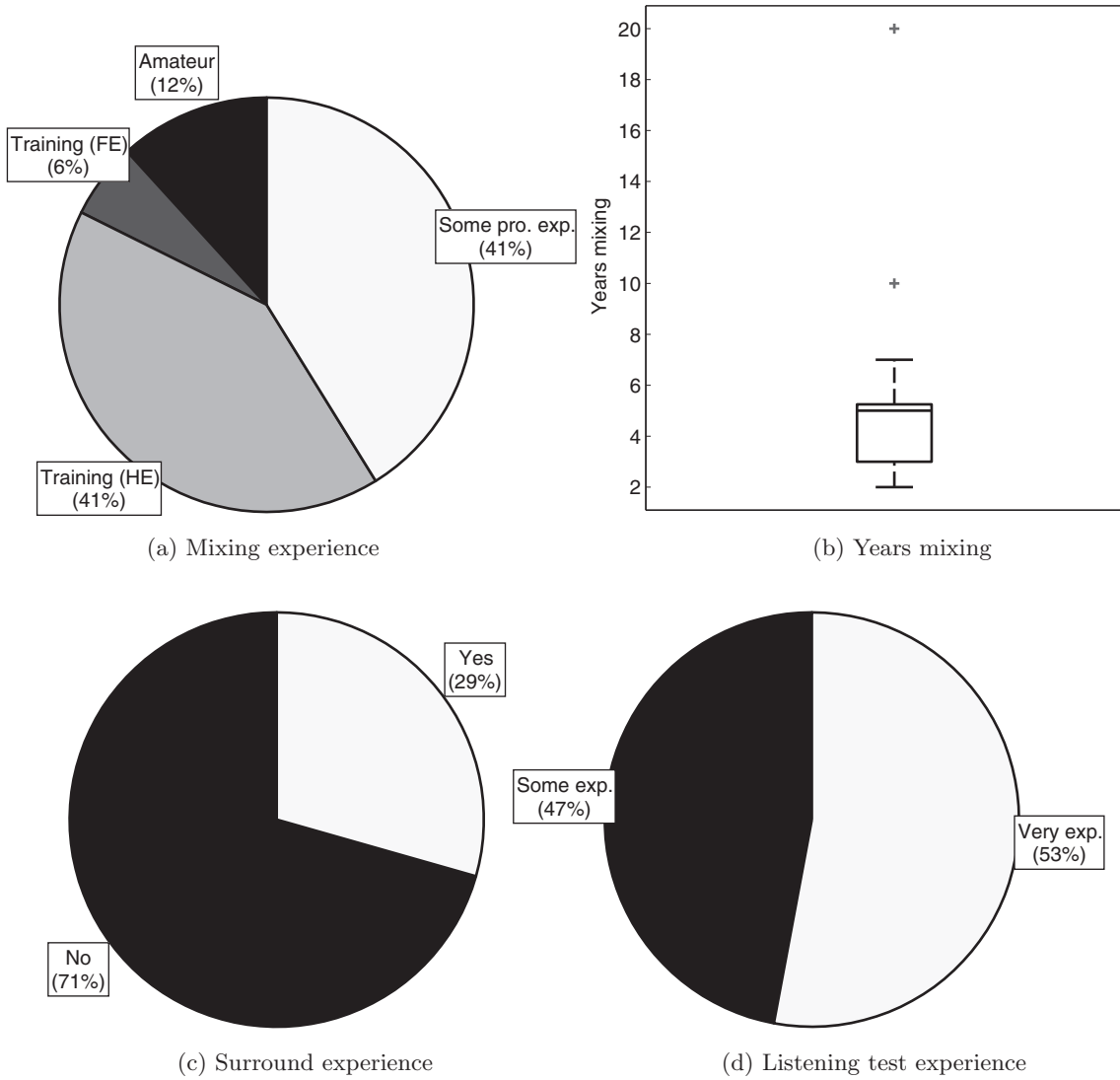
In Table 2, the object categories contained in each program item are detailed.

2.5 Parameters

In addition to the program material, it was necessary to select the parameters that participants would be allowed to vary in order to affect envelopment. The following criteria were used to select the parameters.

1. Parameters that can be varied by metadata changes in the *VISR* (or those that were planned for imminent addition, e.g., reverberation and equalization).
2. Parameters that relate to the perception of envelopment.
3. A total number of parameters that enabled flexibility and control of the mix, while making the test practicable for the participants

The S3A metadata specification was used to determine parameters that met the first criterion. These included: *object level*, *object azimuth*, and *object elevation*. Additionally, current work on the S3A project relating to



(a) Mixing experience

(b) Years mixing

(c) Surround experience

(d) Listening test experience

Fig. 3. Breakdown of participant demographic data

object-based reverberation [34] and equalization meant that these were relevant parameters to consider even though they could not be varied in the renderer at the time of the experiment.

Literature pertaining to envelopment was used to suggest parameters that met the second criterion. An experiment performed by Woodcock et al. [35] investigated the parameters that mixing engineers would use to make manual changes to object-based audio that had been rendered to different loudspeaker layouts, as well as selecting the perceptual attributes that had changed as a result of the rendering. The responses were clustered into a set of parameters (*spread*, *EQ and processing*, *reverb*, *position*, *bass*, and *level*) that were all found to show a significant relationship to changing envelopment. Further controllable parameters found to correlate with envelopment in the lit-

Table 2. Object categories from Woodcock et al.'s [30] experiment that are present in each program item

	Program item			
Category	Radio drama	Football match	Pop track	Jazz Duet
Actions and movement	✓	✓		✓
Continuous background sounds	✓	✓	✓	✓
Transient background sounds	✓	✓	✓	
Clear speech	✓	✓	✓(vocal)	
Non-diegetic music and effects	✓			
Presence of people	✓	✓		
Prominent transient sounds	✓	✓	✓	✓

Table 3. Description of the available parameters, their ranges, and the range of the random offset applied at the start of each test page

Param.	Description	Min.	Max.	Rand. offset range
Level	A boost of cut in the level of the object (or every individual object in a composite object)	−18 dB	+18 dB	±1.5 dB
Azimuth	The horizontal angle of an individual object	−180 deg.	+180 deg.	±5.0 deg.
Elevation	The vertical angle of an individual object	−180 deg.	+180 deg.	±5.0 deg.
Spread	The relative distance between objects in a composite object. A linear multiplier applied to the azimuth and elevation to reposition the object in the range 0 ≤ [new position] ≤ [original position]. The spread control works in the same way for azimuth and elevation of objects.	0	1	±0.1
LF level	A boost or cut in a low-frequency shelf filter (200 Hz, q = 1.0) applied to every object	−18 dB	+18 dB	±1.0 dB
HF level	A boost or cut in a high-frequency shelf filter (2000 Hz, q = 1.0) applied to every object	−18 dB	+18 dB	±1.0 dB
Overall level	A boost or cut in the level of every object	−18 dB	+18 dB	±1.0 dB

erature included detailed parameterization of reverberation (e.g., early decay time, center time, lateral energy fraction, spectral factors, and spatial factors), spectral content, and area of sound distribution [36–39, 17, 14].

The third criterion was assessed based on the results of a pilot test using a similar methodology that was performed with 14 parameters. Qualitative feedback was elicited and analyzed, suggesting that this was the upper limit of parameters that participants were comfortable with. The feedback also showed that there were no clearly missing parameters; some participants requested more detailed control of parameters, but it was not possible to achieve this without making the task too complicated to perform.

It should be noted that there is always a trade-off during selection of parameters for perceptual testing between designing an experiment that is practicable for participants to complete and produces a data set that can be analyzed in a useful manner, and producing results that are as widely generalizable as possible. In this case, some parameters (including the spread or diffuseness of individual objects and movement of objects) were excluded, and parameters such as reverberation that could be varied in a vast number of different ways were simplified. More fine-grained control of such parameters could be the subject of further research.

The parameters (which met all three criteria) that were made available to the listeners, and their ranges, are detailed in Table 3. At the beginning of each test page the parameters were set to the mid-points of their ranges, and a small random offset was applied (as detailed in Table 3). The parameters were used in different combinations depending on the content of each program item; the parameters for each program item are detailed in Table 4. The audio objects in each scene were either treated as individual objects (for example, the narrator in the radio drama scene) or as composite objects (for example, the non-diegetic music in the radio drama scene, which constituted a number of different individual objects).

Limits were applied to the azimuth and elevation of individual objects (therefore also affecting the spread control) depending on the reproduction system. For the 2-channel system, no object's azimuth could lie outside ±30 degrees, and for the 2- and 5-channel systems, no object's elevation could differ from 0 degrees. These limits were applied in metadata processing and could therefore be violated in the raw results; any such deviations were corrected in data preprocessing (Sec. 3.1).

3 ANALYSIS OF OBJECT-BASED PARAMETERS

The method of adjustment experiment described above produced a large amount of data; 17 participants made 36 mixes each, giving a total of 612 mixes. For each mix there were between 8 and 13 parameter values (depending on program item). In the following sections, the results are analyzed in order to address the first research aim (outlined in Sec. 1.3)—determining the relationship between parameters of an object-based mix and the perception of envelopment.

3.1 Data Preprocessing

As mentioned in Sec. 2.5, the limitations on object positions determined by the reproduction system were applied at the metadata processing stage and not in the output from the user control. Therefore, these limits were reapplied to the results.

For the azimuth parameters, it was considered that positioning an object on the left or right side of the sound field was an arbitrary choice, and that the interesting aspect of this choice was the distance from the center. Therefore, the absolute value of the azimuth parameters (originally from −180 degrees to 180 degrees) was taken.

3.2 Participant Clustering

In order to develop an object-based model of envelopment, it is necessary to investigate the relationship between parameters of an object-based mix and the level of envelopment produced. However, it is likely that not all participants in the experiment used the same parameter changes to

Table 4. Parameters in each program material item

Jazz duet	
Object	**Params.**
Piano	Level
	Azimuth
Bass	Level
	Azimuth
Front reverb	Level
Rear reverb	Level
High front reverb	Level
High rear reverb	Level
Overall	Level
	LF level
	HF level

Radio drama	
Object	**Params.**
Narrator	Level
	Azimuth
	Elevation
Music	Level
	Spread
	Reverb level
Sound effects	Level
	Spread
Atmosphere	Level
	Spread
Overall	Level
	LF level
	HF level

Pop track	
Object	**Params.**
Lead vocal	Level
	Azimuth
	Elevation
Flute	Level
	Azimuth
	Elevation
Background	Level
	Spread
Reverb	Level
Overall	Level
	LF level
	HF level

Football match	
Object	**Params.**
Commentator	Level
	Azimuth
	Elevation
Background	Level
	Spread
Overall	Level
	LF level
	HF level

produce each envelopment level; it is possible that different participants may have used different strategies and that some participants were better at performing the task than others. This conjecture is supported by analysis of the distributions of the parameters (396 in total for the four program items, three envelopment levels, and three reproduction methods); 52.53% of the distributions were non-normal (at $p > 0.05$, according to Lilliefors tests performed on the results for each parameter).

Therefore, a clustering analysis was performed to find outlying participants or groups with different strategies. Agglomerative hierarchical clustering with the "average" linkage method was used to generate the clusters [40]. Each participant was initialized as a cluster, and new clusters were formed based on the Euclidean distance between the parameter vectors. The Euclidean distance is sensitive to the absolute values of the parameters, giving more weight to parameters with a higher maximum value. For example, the spread parameters range from 0 to 1, while the level parameters range from −18 to 18 dB. Consequently, the clustering was performed on scaled results; all values were scaled to the range 0–1. For example, for the level parameter, a value of −18 dB was coded as 0, a value of 0 dB was coded as 0.5, and a value of +9 dB was coded as 0.75.

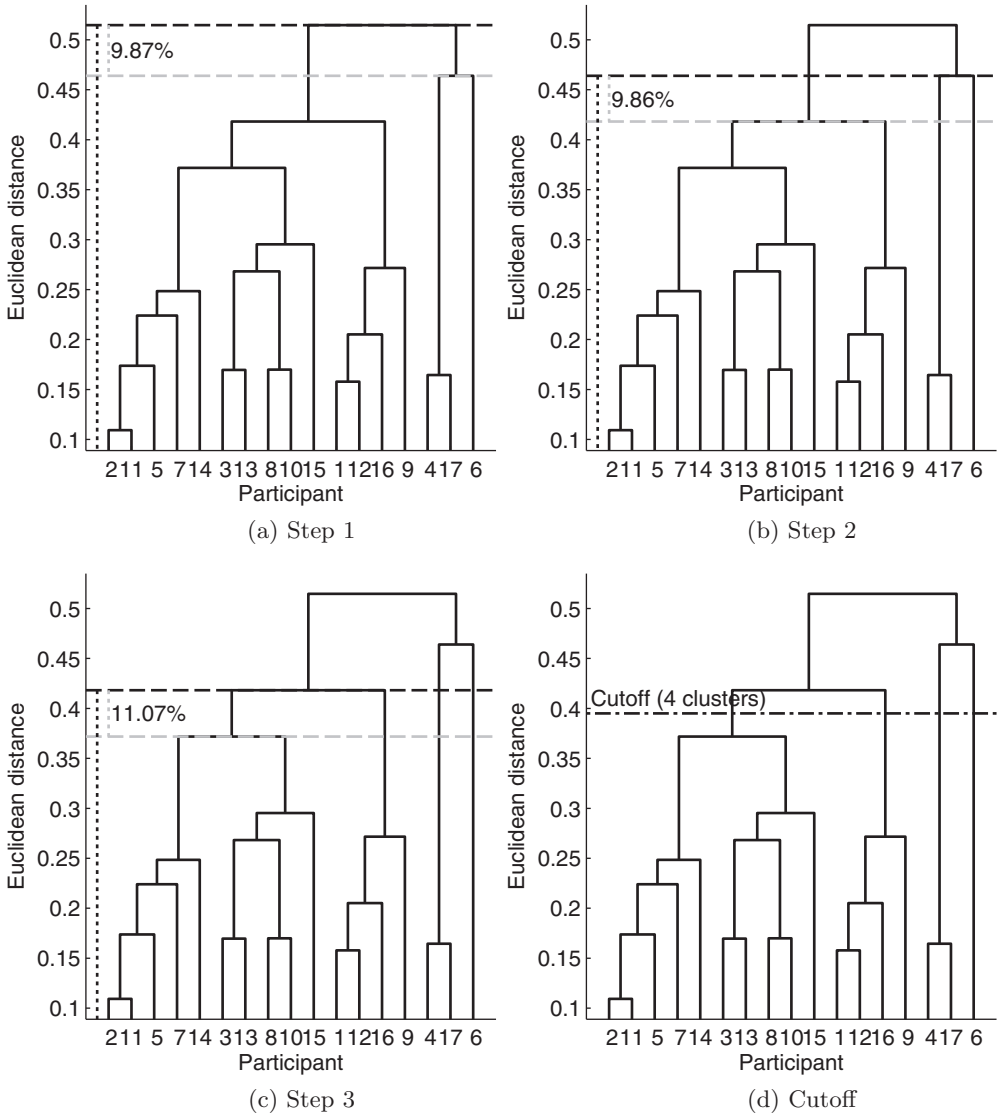
To determine the number of clusters, a cutoff point at a specified inter-cluster distance was calculated. The cutoff distance was determined using the following procedure.

1. Find the distance at the first link (D_1).
2. Find the distance between the first and second links (D_2).
3. If the difference between these two distances is greater than 10% of the first link ($D_1 - D_2 > 0.1{*}D_1$), set the cutoff distance C to the mid-point of the two links (i.e., at $\frac{(D_1+D_2)}{2}$).
4. Otherwise, repeat the procedure for the second and third links, and so on, until the threshold is reached. If the threshold is never reached, each cluster will feature one participant.

The 10% threshold was determined by observing the dendrograms and ensuring that this value led to sensible cuts (i.e., those that a human interpreter would be likely to make, as this is commonly the way in which clustering solutions are interpreted). An example of the steps for this procedure is shown in Fig. 4.

This clustering method was used for all program items, reproduction methods, and envelopment levels. This resulted in clustering solutions with two to five clusters. In the majority of cases (72%), two clusters were found; observation of the dendrograms suggested that in most cases, the clustering algorithm was generally finding a small number of outlying participants rather than separate groups with clearly distinct strategies. This is supported by the fact that on average, 84% of participants fell into the largest cluster. A dendrogram exemplifying this trend is shown in Fig. 5a. However, in a minority of cases, two distinct groups did seem to form (see, for example, Fig. 5b).

In most cases one or more small outlying clusters were found; therefore, an assessment of the participants that fell outside of the main cluster was performed. Fig. 6 shows the percentage of cases in which each participant fell outside of the largest cluster. It is notable that no participant fell into the largest cluster in every case; rather, all participants fell outside of the largest cluster in approximately 5–20% of cases, with the exception of participants 16 and 17, who lay outside the main cluster slightly more frequently



(a) Step 1 (b) Step 2

(c) Step 3 (d) Cutoff

Fig. 4. Stepwise method for determining clustering distance cutoff (for the football match program item, 5-channel reproduction, low envelopment)

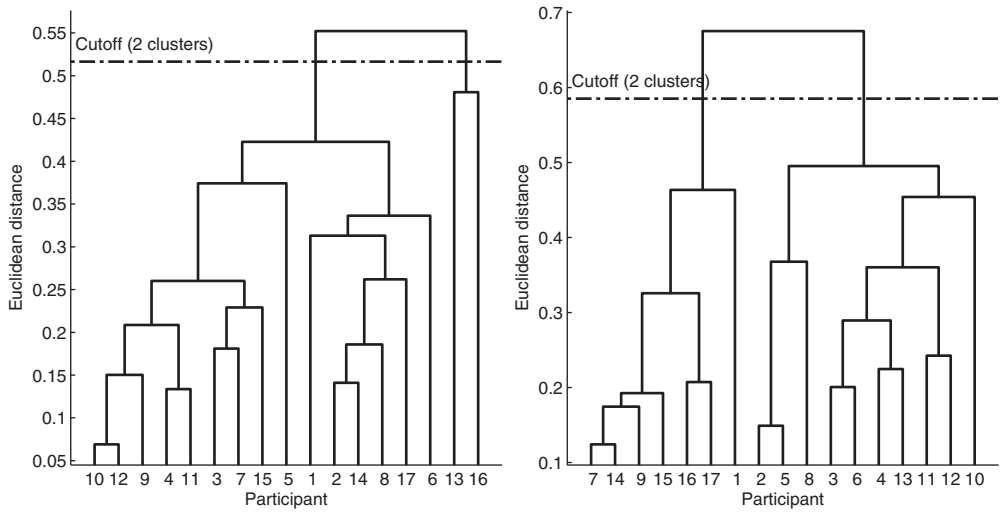
(approximately 30% and 40% of cases respectively). There was found to be no significant relationship between this percentage and any of the demographic variables described in Sec. 2.2.

As a result of the analysis presented above, all participants were included in subsequent analysis; however, only parameter values from the largest cluster were included. In future work it may be of interest to compare across clusters (for example, collecting ratings of perceived similarity, mix quality, or envelopment), but this is not attempted here.

3.3 Analysis of Mix Parameters

The primary aim of the experiment described in this paper was to establish the relationships between parameters of object-based audio mixes and the perceived level of envelopment. These relationships were analyzed for different levels of envelopment (low, medium, and high) and for the three reproduction systems under test (2-channel stereo, 5-channel surround sound, and 22-channel surround sound). It should be noted that the envelopment that can be achieved in a 22-channel setup is likely to be much greater than can be achieved in a 2-channel setup. The categorical data available here precludes analysis of the absolute level of envelopment produced in the different systems but is suitable for analysis of the parameter settings for producing a range of envelopment levels for each system.

The parameters made available to participants were different for each of the program material items; consequently, each item was analyzed separately (Sec. 3.3.1). However, it is also possible to group the parameters across program items, which is important for developing a more generalizable predictive model or envelopment modification tool. The object types determined by Woodcock et al. [30] were used as the basis for this parameter grouping; for example, the level for the “clear speech” group is based on the narrator level from the radio drama program material, the



(a) Small outlying group (b) Two distinct groups

Fig. 5. Example dendrograms showing a small outlying group and two distinct groups (pop song, 5-channel reproduction, medium and high envelopment respectively)

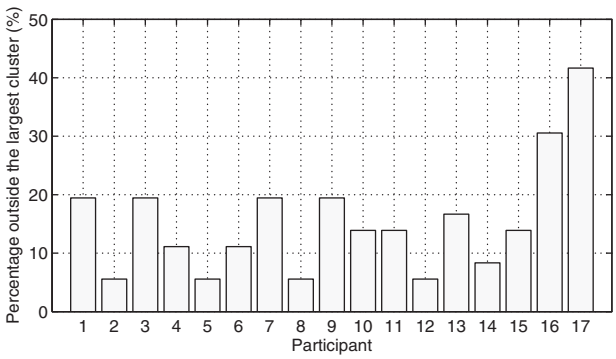


Fig. 6. Percentage of cases in which each particiapant fell outside of the largest cluster

lead vocal level from the pop track, and the commentator level from the football match. This analysis is discussed in Sec. 3.3.2.

As discussed in Sec. 3.2 all analysis below is based on the largest participant cluster for each mix.

3.3.1 Mix Parameters for Each Program Item

Fig. 7 shows mean parameter values (with 95% confidence intervals determined using the *t*-distribution) for the low, medium, and high envelopment mixes and the three reproduction systems. For each parameter and reproduction method, a multinomial logistic regression model was produced to test the significance (at $\alpha = 1\%$) of the effect of the parameter on the envelopment level (i.e., testing whether a change in the parameter significantly affects the odds that a mix has medium rather than low envelopment, or high rather than medium envelopment).

In the majority of cases, the various mix controls were used in a similar manner across the three reproduction methods, with a small number of exceptions. For the jazz duet program item, the rear reverberation parameters were generally set lower for the 2-channel mix (presumably as the reverberation was folded into the front channels and this was an attempt to keep the direct-to-reverberant ratio at an acceptable level). Also for the jazz duet item, the piano and bass azimuth modifications were more pronounced for the 22-channel mix. The only other apparent deviations between reproduction formats appeared in the cases where the participants were limited by the reproduction format.

In general, the object level controls were not found to be significant; they were only significant when used with reverberation objects. The low frequency level was significant for every program item (as low frequency level increased, envelopment increased).

The spread controls showed the most pronounced relationship with the three envelopment levels (with a significant relationship in every case): spread was increased for the higher levels of envelopment.

3.3.2 Mix Parameters across Object Types

In order to develop generalizable rules for modification of envelopment in object-based mixes, it is necessary to consolidate results across program items. The parameters were grouped based on the object categories discussed in Sec. 2.4 [30] (with slight modifications due to the relatively low resolution of the group objects in this experiment) and the parameters detailed in Table 3. This resulted in 15 group parameters, detailed in Table 5.

The mean value of each group parameter for the three reproduction systems is given in Table 6. Fig. 8 shows the mean group parameter values (with 95% confidence intervals determined using the *t*-distribution) for the low, medium, and high envelopment mixes and the three reproduction systems.

As was seen in Sec. 3.3.1, the group parameter values were consistent across reproduction methods. Logistic regression models were produced in the same manner as described in Sec. 3.3.1. The group parameters

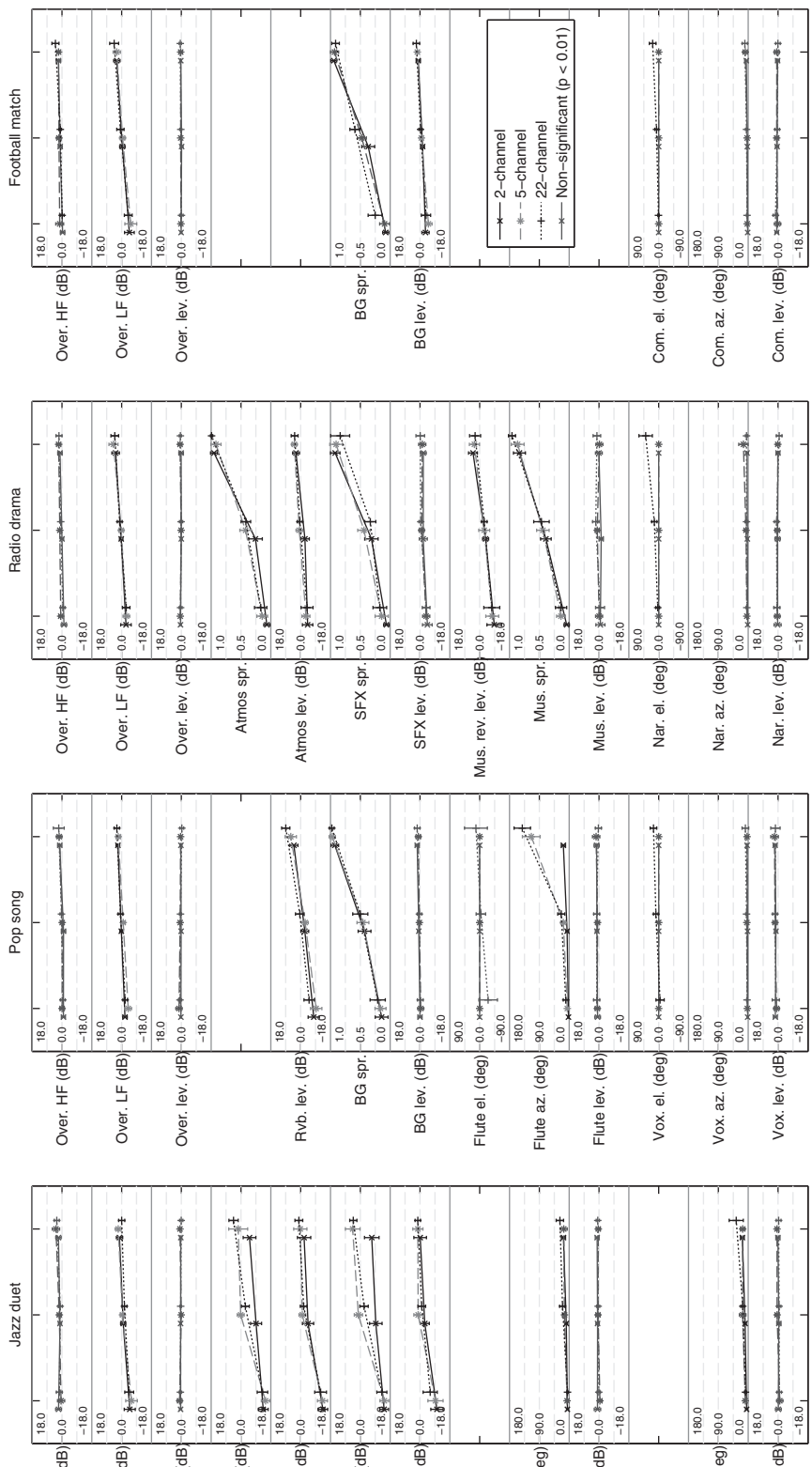


Fig. 7. Parameter values for the low, medium, and high envelopment mixes. Reproduction methods are indicated by shade, line style, and marker style; program items in panels. Solid light gray lines indicate parameters that were non-significant in a logistic regression model (see body text for further explanation). The parameters are approximately vertically aligned so that similar parameters are on the same level.

Table 5. Group parameters, including the program items and objects or group objects that contributed to the values for each parameter.

	Contributing program items			
Group parameter	Jazz	Pop	Radio	Football
Overall level	✓	✓	✓	✓
Overall HF level				
Overall LF level				
Clear speech level		✓Vocal	✓Narrator	✓Commentator
Clear speech azimuth				
Clear speech elevation				
Foreground sounds level	✓Piano, bass	✓Flute	✓SFX	
Foreground sounds azimuth	✓Piano, bass	✓Flute		
Foreground sounds elevation		✓Flute		
Foreground sounds spread			✓SFX	
Background sounds level		✓Background	✓Atmos	✓Background
Background sounds spread				
Non-diegetic music and effects level			✓Music	
Non-diegetic music and effects				
Reverberation level	✓Front, back, high front, high back	✓Reverb.	✓Music reverb.	

that showed a statistically significant relationship to envelopment level were: overall HF level, overall LF level, clear speech elevation, foreground sounds azimuth, foreground sounds spread, background sounds level, background sounds spread, non-diegetic music spread, and reverberation level. All spread parameters showed a significant relationship with the level of envelopment produced.

3.3.3 A Logistic Regression Model of Envelopment in Object-Based Audio Mixes

Multinomial logistic regression can be used to predict category membership in a response variable from a number of independent variables [41]. In this case, a model was trained to predict whether a mix would fall into the low, medium, or high envelopment categories, using the object-based metadata (i.e., the group parameters introduced in Sec. 3.3.2). When the group parameters were calculated, data from participants falling outside of the largest cluster (see Sec. 3.2) were omitted. However, this results in an unbalanced dataset (i.e., there are a different number of observations for different group parameters). In order to train the logistic regression model with a balanced dataset, values that were omitted due to the clustering were replaced by the mean of the available data.

Similarly, the dataset was unbalanced because the different group parameters comprised different numbers of parameters; for example, overall level was made up of four parameters (for the four program items), while foreground sounds spread only included one parameter (SFX spread from the radio drama). To facilitate production of a model, the mean was taken across each of the individual parameters that made up the group parameters.

Before fitting the logistic regression model, the variance inflation factors (VIFs) were calculated for the group parameters. VIFs quantify the multicollinearity in a feature set, i.e., the degree to which the features are linearly related; where there is such a relationship between

Table 6. Mean values for the significant group parameters at low, medium, and high envelopment. Significant parameters are indicated by an asterisk (*).

	2-channel			5-channel			22-channel		
Group parameter	Low	Med	High	Low	Med	High	Low	Med	High
Ov. lev. (dB)	0.31	0.08	0.24	0.86	0.30	0.58	0.41	0.33	0.27
Ov. HF lev. (dB)*	0.00	0.35	1.68	0.54	1.11	2.30	0.10	0.76	2.87
Ov. LF lev. (dB)*	−3.58	−0.29	2.51	−4.46	−0.44	3.15	−3.38	0.16	2.70
Clear speech lev. (dB)	0.99	0.76	0.71	0.60	1.03	0.86	1.20	1.11	0.19
Clear speech az. (deg)	1.71	2.50	4.25	3.24	4.78	9.81	3.82	5.17	8.20
Clear speech el. (deg)*	-	-	-	-	-	-	−0.14	9.24	24.62
Fg. sounds lev. (dB)	−0.37	0.02	0.17	−1.02	0.22	0.19	−0.70	0.26	0.16
Fg. sounds az. (deg)*	4.13	8.24	17.51	6.85	16.35	45.24	7.88	20.10	53.99
Fg. sounds el. (deg)	-	-	-	-	-	-	−25.50	−3.63	11.38
Fg. sounds spread*	0.06	0.31	0.92	0.14	0.44	0.91	0.17	0.33	0.84
Bg. sounds lev. (dB)*	−2.23	−1.03	1.84	−2.58	0.11	2.32	−2.55	0.15	2.54
Bg. sounds spread*	0.10	0.36	0.94	0.14	0.45	0.94	0.21	0.51	0.95
Non-diegetic mus. lev. (dB)	−0.83	−1.32	0.10	−0.03	1.07	−0.21	−0.62	1.96	1.29
Non-diegetic mus. spread*	0.04	0.39	0.83	0.14	0.44	0.86	0.14	0.46	0.96
Reverb lev. (dB)*	−11.51	−5.42	−1.02	−11.35	−0.86	2.76	−9.43	−1.64	3.38

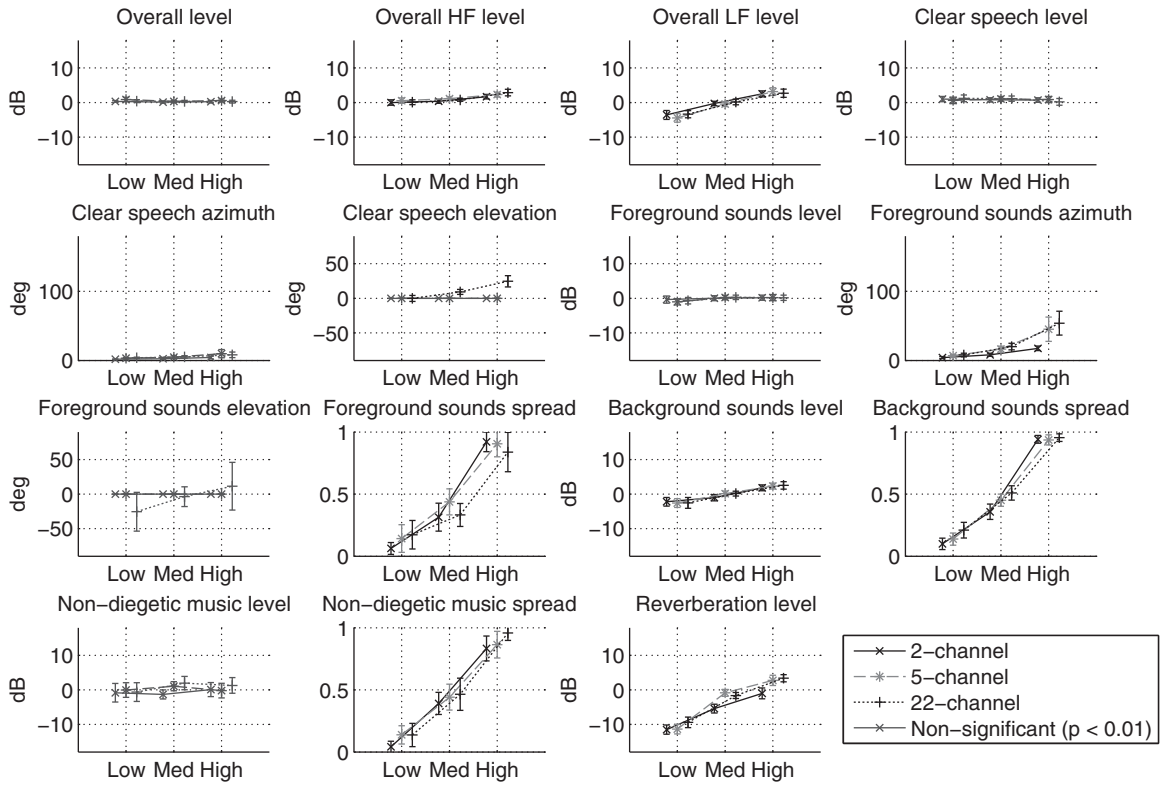


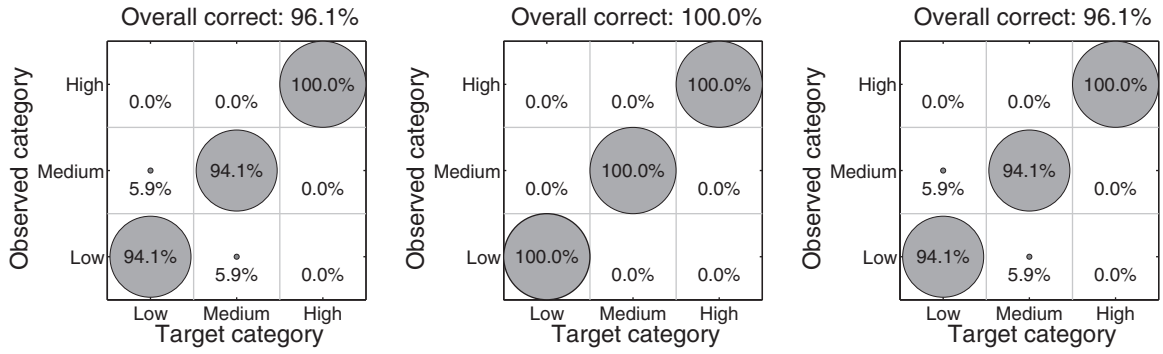
Fig. 8. Group parameter values for the low, medium, and high envelopment mixes. Low, medium, and high envelopment mixes are shown on the *x*-axis. Reproduction methods are indicated by shade, line style, and marker style. Solid light gray lines indicate parameters that were non-significant in a logistic regression model (see body text for further explanation).

features, the model parameter estimates become unreliable [41]. A maximum VIF (VIF_{max}) greater than 10 or a mean VIF (VIF_{mean}) greater than 1 are indications that features are unacceptably related [42, 43]. In this case, VIF_{max} was 22.78, 19.92, and 14.41 for 2-channel, 5-channel, and 22-channel respectively. VIF_{mean} was 5.73, 5.09, and 4.94 for 2-channel, 5-channel, and 22-channel respectively. These values suggest intolerably high multicollinearity. This is unsurprising given the nature of the relationships shown in Fig. 8, which show increasing values for higher envelopment levels for all significant features. Consequently, an overall logistic regression model was not produced.

In order to remove the correlation between parameters, a principal component analysis (PCA) was performed in order to determine the quantitative relationships between features. Analysis of a scree plot showed a knee-point at the first component, which accounted for approximately 70% of the variance in feature values (for all three reproduction methods). Consequently, only a single principal component was used in the models. A *k*-fold cross-validation procedure was then used to train and evaluate logistic regression models for each reproduction system; the following procedure was used.

1. From the feature vector, randomly designate 20 data points as test cases and the remaining (31) points as training cases
2. Calculate *z*-scores for the training features, also saving the mean and standard deviation
3. Perform PCA on the training set *z*-scores
4. Perform logistic regression on the first principal component
5. Standardize the test set according to the mean and standard deviation from the training set, and then calculate the loadings onto the principal components (using the coefficients calculated for the training set)
6. Use the logistic regression model to predict category membership for the PCA solution of the test set
7. Calculate the percentage of correct classifications
8. Repeat steps 1–7 for 500 iterations
9. Calculate the mean percentage of correct classifications across all iterations

The resulting mean percentages of correct classification were 95.5%, 99.3%, and 94.6% for 2-, 5-, and 22-channel reproduction respectively. This compares favorably to the values without cross-validation of 96.1%, 100.0%, and 96.1% respectively, suggesting that the model is likely to generalize well to data from outside of the training set. Fig. 9 shows confusion matrix plots for the full model (i.e., without cross-validation). It can be seen that no large misclassifications are made (i.e., in no cases is a low envelopment mix classified as high envelopment, or a high envelopment mix classified as low envelopment).



(a) 2-channel (b) 5-channel (c) 22-channel

Fig. 9. Confusion matrices for PCA models

3.4 Summary

In Sec. 3 the results from the method of adjustment experiment described in Sec. 2 were analyzed in order to determine the relationship between parameters of an object-based mix and the perception of envelopment. A clustering analysis was performed, showing that there were some differences between participants, but that these were limited to small outlying groups rather than clearly defined strategies. Consequently, results from all participants were used in the analysis. The parameter values were for the most part consistent across the different reproduction methods, with only minor deviations observed. The spread parameters showed the most pronounced relationship with envelopment level. When parameters were combined in order to analyze results across different program items, nine parameters showed statistically significant relationships to envelopment level: overall HF level, overall LF level, clear speech elevation, foreground sounds azimuth, foregrounds sounds spread, background sounds level, background sounds spread, non-diegetic music spread, and reverberation level. A logistic regression model was trained and found to be able to accurately predict envelopment category membership, suggesting a strong relationship between the parameters and the level of envelopment produced. However, it was also determined that there was a high degree of multicollinearity between the parameter values. It is not clear how envelopment would be affected were the parameters to vary with less multicollinearity; however, this result suggests that envelopment could potentially be modified in the absence of one parameter by modification of another (i.e., if horizontal spread of sound could not be increased, an increase in reverberation level would also have the effect of increasing envelopment).

4 VALIDATION OF PARAMETERS FOR ENVELOPMENT MODIFICATION

The experiment reported above was used to determine a set of parameters that affect perceived envelopment. The second aim outlined in Sec. 1.3 is to develop and test a system for manipulating envelopment in object-based audio in a perceptually relevant manner. Consequently, the parameters determined in Sec. 3 were used to generate different levels of envelopment in a set of test mixes. To validate the performance of the system, and consequently the applicability of the parameters and their values, the level of envelopment produced was evaluated using an existing envelopment model.

4.1 Program Items

Four program items were used. Two of the items were excerpts from the same content as used in the training set described in Sec. 2.4 (the radio drama forest scene and the pop track); however, different excerpts were used. The remaining two items—detailed below—were not part of the original training set.

- Radio drama scene (the *Protest* scene the from S3A object-based audio drama dataset [31]), featuring speech, foreground sounds, background sounds, non-diegetic music, and reverberation.
- Live rock music recording (with vocals, foreground sounds, background sounds, and reverberation).

Each excerpt was 20 seconds long.

4.2 Methodology

The significant parameters from Table 6 were used to modify the relevant parameters in the object-based mixes described above, with the exception of "foreground sounds azimuth," which is not suitable for this type of general processing as it refers to the position of single predetermined objects in the mix, and "clear speech elevation," as a 5-channel system was used (as discussed below). Therefore, a total of seven group parameters were varied.

The group parameter values were set by applying a piecewise linear mapping to the values in Table 6, as shown in Eq. (1) (where V_g is the new value for group parameter g, E is the target envelopment in the range [0, 100], and $P_{g_{\text{low}}}$, $P_{g_{\text{med}}}$, and $P_{g_{\text{high}}}$ are the mean parameter values for the gth group parameter at low, medium, and high envelopment respectively).

$$V_g = \begin{cases} \left(\frac{P_{g_{\text{med}}} - P_{g_{\text{low}}}}{50} \times E\right) + P_{g_{\text{low}}}, & \text{if } E <= 50 \\ \left(\frac{P_{g_{\text{high}}} - P_{g_{\text{med}}}}{50} \times E\right) + P_{g_{\text{med}}}, & \text{if } E > 50. \end{cases} \quad (1)$$

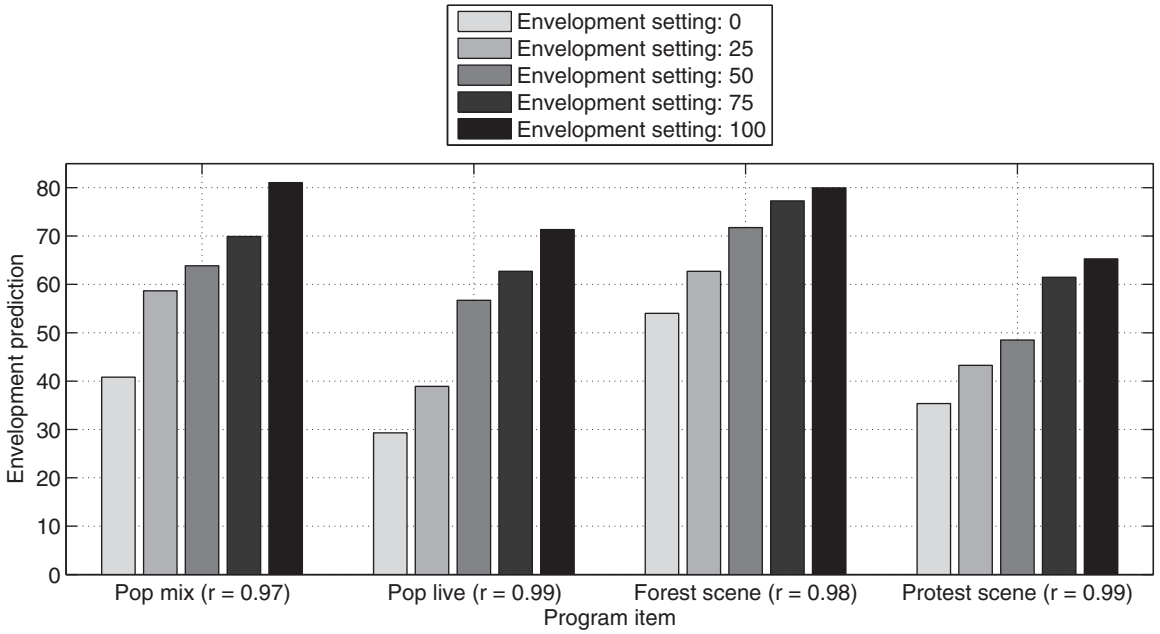


Fig. 10. Envelopment predictions for four validation mixes at five envelopment levels. All specified values of r (Pearson's correlation coefficient) are significant at $p < 0.01$.

The target envelopment E was set to 0, 25, 50, 75, and 100. The program items were produced by setting the parameters in the *Metadapter*, rendering to a 5-channel setup, and capturing the loudspeaker feeds. Envelopment was predicted using an implementation of George et al.'s [14] model. The 5-channel rendering was used as this model is designed for 5-channel content. George et al.'s [14] model was selected as it showed a good fit to its training data (RMSE $= 8.54\%$, $R = 0.90$) and the relevant features can be extracted more simply from a 5-channel signal than those used by Conetta [26] or Dewhirst [23]. The model takes loudspeaker feeds as input; predictions are based on a linear combination of features relating to spectral content, IACC (calculated from a binaural simulation at a variety of head angles), inter-channel coherence (calculated from a Karhunen-Loève transform (KLT) decomposition of the input channels), and angle of arrival of sound (also calculated from the KLT signals).

4.3 Results

Fig. 10 shows the envelopment predictions for each program item and target envelopment level. The figure shows that, in all cases, the envelopment predictions monotonically increase with the target envelopment, suggesting that the parameters determined are suitable for manipulating envelopment in a range of stimuli (including those outside of the original training data set). There is a high positive correlation between envelopment predictions and target envelopment for all program items ($r > 0.97$, $p < 0.01$ in all cases), suggesting that the linearity of the relationship is strong. However, this result should be interpreted in the light of the small number of data points; it is clear that there are deviations from a linear relationship, for example, in the case of the radio drama scenes, for which the 75 and 100 target envelopment settings were given similar envelopment scores by the model.

4.4 Summary

In Sec. 4, the parameters determined in Sec. 3 were used to vary the level of envelopment in new stimuli in order to test a system for manipulating envelopment in object-based audio. The results showed that the parameters can be used to change the level of envelopment in a perceptually relevant manner, showing a monotonic relationship between target and predicted envelopment. However, further work is required to assess the detailed relationships between target and perceived envelopment.

5 DISCUSSION AND CONCLUSIONS

There were two primary aims of the work reported above: (i) to determine the relationship between parameters of an object-based mix and the perception of envelopment; and (ii) to develop and test a system for manipulating envelopment in object-based audio in a perceptually relevant manner.

A method of adjustment experiment was performed in which mixing engineers were asked to create mixes of object-based content at three levels of envelopment (low, medium, and high) while keeping the overall mix quality at an acceptable level. This enabled analysis of parameter values in order to assess how participants created different levels of envelopment. A clustering analysis was performed to see if there were different strategies employed for creating envelopment. The results suggested that there were not clearly defined groups; however, outlying mixes were excluded from the results on a case-by-case basis.

The parameters investigated were high-level features that can be varied in object-based audio (as metadata changes), based on levels, positions, and equalization of objects or groups of objects. Many of the parameters showed significant relationships with envelopment level. In particular, the low frequency level, reverberation level, and spread parameters were always significant. High frequency level was only significant for one program item and overall level was not used to vary envelopment. In order to compare across program items, the parameters were grouped. Of the group parameters, the spread of composite objects (i.e., the relative positions of individual objects making up a composite object—as described in Table 3) was found to be particularly important. The significant parameters were found to be highly correlated; as parameter values increased, so did the level of envelopment. Consequently, it proved difficult to assess the individual contribution of each parameter.

There were found to be few differences in parameter settings between the different reproduction methods. However, the high envelopment category for 2-channel reproduction will naturally be less enveloping than the high envelopment category for 22-channel reproduction in an absolute sense. This requires further investigation (as discussed in Sec. 5.1).

The parameters that were investigated involve relative changes to the objects in a mix; consequently, the experiments reported above do not allow prediction of envelopment from metadata or from the produced sound field. It would be interesting to assess parameters of the produced sound field—direct-to-reverberant ratios, the perceived loudnesses of objects, speech-to-background ratios, and so on—to ascertain how such signal-level features affect envelopment and how they could be used to predict perceived envelopment.

The experiment described in Sec. 4 demonstrated that the parameters determined can be used to change envelopment. The parameter values determined above were used to create mixes at a range of target envelopment levels. An implementation of George et al.'s [14] model was used to predict the envelopment achieved; the predictions were found to correlate strongly with the target envelopment.

Envelopment has been shown to be one of the most important attributes of listener preference for spatial audio reproduction. Some potential applications for the envelopment modification system presented in this paper include development of tools to help producers create enveloping mixes or to create mixes at different envelopment levels. The envelopment modification tool could also be used to optimize or personalize envelopment in audio reproduction in the home.

5.1 Future Work

There are a number of avenues for further exploration of envelopment in channel- or object-based audio. It would be interesting to quantitatively assess the envelopment level that was produced by individual participants and groups of participants. Such data could be used to assess the ability of participants with different levels of experience to create a target level of envelopment, and to test, train, and evaluate existing or new models of envelopment (including those reviewed in Sec. 1.2) with more detail than the coarse categories used in this paper.

In this work three reproduction systems were used; low, medium, and high envelopment level mixes were produced in each case. It would be interesting to consider the absolute levels of envelopment that were able to be produced in each system. For example, it is likely that the high envelopment produced in the 2-channel system is lower in absolute terms than the high envelopment produced in the 22-channel system.

It would also be beneficial to look at some of the parameters in more detail. For example, reverberation level was investigated here, but the literature review suggested that specific aspects of reverberation contribute to the perception of envelopment. Object-based reverberation [34] provides a good opportunity for investigating the effects of reverberation on envelopment in detail. The spread control was shown to be important, but more detail could be collected on how absolute object positions influence envelopment (in different types of scene and for different categories of object).

ACKNOWLEDGMENTS

This work was supported by the EPSRC program Grant S3A: Future Spatial Audio for an Immersive Listener Experience at Home (EP/L000539/1) and the BBC as part of the BBC Audio Research Partnership.

Details about the data underlying this work, along with the terms for data access, are available from https://doi.org/10.15126/surreydata.00845727.

The authors would like to thank Andreas Franck for providing support for the *VISR Renderer* and *Metadapter* software and Rupert Flindt and Craig Cieciura for producing the live rock music item used in the validation experiment.

REFERENCES

[1] G. Thomas, A. Engström, J.-F. Macq, O. A. Niamut, B. G. Shirley, and R. Salmon, "State-of-the-Art and Challenges in Media Production, Broadcast and Delivery," in O. Schreer, J.-F. Macq, O. A. Niamut, J. Ruiz-Hidalgo, B. Shirley, G. Thallinger, G. Thomas (eds.), *Media Production, Delivery and Interaction for Platform Independent Systems*, pp. 5–73 (John Wiley & Sons, Ltd., 2013).

[2] H. Stenzel and U. Scuda, "Producing Interactive Immersive Sound for MPEG-H: A Field Test for Sports Broadcasting," presented at the *137th Convention of the Audio Engineering Society* (2014 Oct.), convention paper 9211.

[3] R. G. Oldfield, B. G. Shirley, and J. Spille, "Object-Based Audio for Interactive Football Broadcast," *Multimedia Tools and Applications*, vol. 74, pp. 2717–2741 (2015), https://doi.org/10.1007/s11042-013-1472-2.

[4] C. Pike, R. Taylor, T. Parnell, and F. Melchior, "Object-Based 3D Audio Production for Virtual Reality Using the Audio Definition Model," presented at the *AES*

International Conference: Audio for Virtual and Augmented Reality (2016 Sep.), conference paper 2-1.

[5] 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.

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

[7] U. Scuda, H. Stenzel, and D. Baxter, "Using Audio Objects and Spatial Audio in Sports Broadcasting," presented at the *AES 57th International Conference: The Future of Audio Entertainment Technology—Cinema, Television and the Internet* (2015 Mar.), conference paper 5-2.

[8] J.-M. Jot, B. Smith, and J. Thompson, "Dialog Control and Enhancement in Object-Based Audio Systems," presented at the *139th Convention of the Audio Engineering Society* (2015 Oct.), convention paper 9356.

[9] J. Francombe, T. Brookes, and R. Mason, "Perceptual Evaluation of Spatial Audio: Where Next?" *Proceedings of the 22nd International Congress on Sound and Vibration, Florence, Italy, 12–16 July* (2015).

[10] 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.

[11] 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, pp. 968–976 (2005), https://doi.org/10.1121/1.1945368.

[12] J. S. Bradley and G. A. Soulodre, "Listener Envelopment: an Essential Part of Good Concert Hall Acoustics," *J. Acoust. Soc. Amer.*, vol. 99, pp. 22–22 (1996), https://doi.org/10.1121/1.414533.

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

[14] S. George, S. Zielinski, F. Rumsey, P. J. B. Jackson, R. Conetta, M. Dewhirst, D. Meares, and S. Bech, "Development and Validation of an Unintrusive Model for Predicting the Sensation of Envelopment Arising from Surround Sound Recordings," *J. Audio Eng. Soc.*, vol. 58, pp. 1013–1031 (2010 Dec.).

[15] F. Rumsey, "Spatial Quality Evaluation for Reproduced Sound: Terminology, Meaning, and a Scene-Based Paradigm," *J. Audio Eng. Soc.*, vol. 50, pp. 651–666 (2002 Sep.).

[16] D. Griesinger, "Objective Measures of Spaciousness and Envelopment," presented at the *AES 16th International Conference: Spatial Sound Reproduction* (1999 Mar.), conference paper 16-003.

[17] G. A. Soulodre, M. C. Lavoie, and S. G. Norcross, "Objective Measures of Listener Envelopment in Multichannel Surround Systems," *J. Audio Eng. Soc.*, vol. 51, pp. 826–840 (2003 Sep.).

[18] M. Morimoto and K. Iida, "A Practical Evaluation Method of Auditory Source Width in Concert Halls," *J. Acoust. Soc. Japan (E)*, vol. 16, pp. 59–69 (1995), https://doi.org/10.1250/ast.16.59.

[19] R. Conetta, S. Zielinski, F. Ramsey, and P. J. B. Jackson, "Envelopment: What is it? A Definition for Multichannel Audio," *1st SPACE-Net Workshop* (University of York, 2007).

[20] 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.

[21] J. van Dorp Schuitman, D. de Vries, and A. Lindau, "Deriving Content-Specific Measures of Room Acoustic Perception Using a Binaural, Nonlinear Auditory Model," *J. Acoust. Soc. Amer.*, vol. 133, pp. 1572–1585 (2013), https://doi.org/10.1121/1.4789357.

[22] J. Nowak and S. Klockgether, "Perception and Prediction of Apparent Source Width and Listener Envelopment in Binaural Spherical Microphone Array Auralizations," *J. Acoust. Soc. Amer.*, vol. 142, pp. 1634–1645 (2017), https://doi.org/10.1121/1.5003917.

[23] M. Dewhirst, *Modelling Perceived Spatial Attributes of Reproduced Sound*, Ph.D. Thesis, Institute of Sound Recording, University of Surrey (2008).

[24] B. Supper, *An Onset-Guided Spatial Analyser for Binaural Audio*, Ph.D. Thesis, Institute of Sound Recording, University of Surrey, Guildford, UK (2005).

[25] P. Power, B. Davies, and J. Hirst, "Investigation into the Impact of 3D Surround Systems on Envelopment," presented at the *137th Convention of the Audio Engineering Society* (2014 Oct.), convention paper 9108.

[26] R. Conetta, *Towards the Automatic Assessment of Spatial Quality in the Reproduced Sound Environment*, Ph.D. Thesis, Institute of Sound Recording, University of Surrey (2010).

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

[28] R. Mason, "Installation of a Flexible 3D Audio Reproduction System into a Standardized Listening Room," presented at the *140th Convention of the Audio Engineering Society* (2016 May), e-Brief 256.

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

[30] 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.

[31] 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 May), e-Brief 255.

[32] R. G. Oldfield, and B. G. Shirley, "Automatic Mixing and Tracking of On-Pitch Football Action for Television Broadcasts," presented at the *130th Convention of the Audio Engineering Society* (2011 May), convention paper 8346.

[33] J. Francombe, T. Brookes, R. Mason, R. Flindt, P. Coleman, Q. Liu, and P. J. B. Jackson, "Production and Reproduction of Programme Material for a Variety of Spatial Audio Formats," presented at the *138th Convention of the Audio Engineering Society* (2015 May), e-Brief 199.

[34] P. Coleman, A. Franck, P. J. B. 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.

[35] J. Woodcock, W. J. Davies, F. Melchior, and T. J. Cox, "Elicitation of Expert Knowledge to Inform Object-Based Audio Rendering to Different Systems," *J. Audio Eng. Soc.*, vol. 66, pp. 43–58 (2018 Jan./Feb.), https://doi.org/10.17743/jaes.2018.0001.

[36] J. S. Bradley and G. A. Soulodre, "The Influence of Late Arriving Energy on Spatial Impression," *J. Acoust. Soc. Amer.*, vol. 97, pp. 2263–2271 (1995), https://doi.org/10.1121/1.411951.

[37] M. Morimoto, and K. Iida, "Effects of Front/back Energy Ratios of Early and Late Reflections on Listener Envelopment," *J. Acoust. Soc. Amer.*, vol. 103, pp. 2748–2748 (1998), https://doi.org/10.1121/1.422799.

[38] J. Blauert and W. Lindemann, "Auditory Spaciousness: Some Further Psychoacoustic Analyses," *J. Acoust. Soc. Amer.*, vol. 80, pp. 533–542 (1986), https://doi.org/10.1121/1.394048.

[39] D. Griesinger, "The Importance of the Direct to Reverberant Ratio in the Perception of Distance, Localization, Clarity, and Envelopment," presented at the *126th Convention of the Audio Engineering Society* (2009 May), convention paper 7724.

[40] B. S. Everitt, S. Landau, M. Leese, and D. Stahl, "Hierarchical Clustering," in *Cluster Analysis*, pp. 71–110 (Wiley, Chichister, 2011), https://doi.org/10.1002/9780470977811.ch4.

[41] A. P. Field, *Discovering Statistics Using SPSS* (Sage Publications, London, 2005), pp. 264–315.

[42] B. L. Bowman and R. T. O'Connell, *Linear Statistical Models: An Applied Approach* (Duxbury Press, Belmont, CA, USA, 1990).

[43] R. H. Myers, *Classical and Modern Regression with Applications* (Duxbury Press, Belmont, 1990).

THE AUTHORS

Jon Francombe

Tim Brookes

Russell Mason

Jon Francombe graduated with a first-class honors 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" 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.

•

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.

•

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 modelling aspects of auditory perception.