Audio Engineering Society

Conference Paper

Presented at the Conference on
Spatial Reproduction
2018 August 6 – 9, Tokyo, Japan

A framework for intelligent metadata adaptation in object-based audio

James Woodcock[1], Jon Francombe[2], Andreas Franck[3], Philip Coleman[4], Richard Hughes[1], Hansung Kim[5], Qingju Liu[5], Dylan Menzies[3], Marcos F. Simón Gálvez[3], Yan Tang[1], Tim Brookes[4], William J. Davies[1], Bruno M. Fazenda[1], Russell Mason[4], Trevor J. Cox[1], Filippo Maria Fazi[3], Philip J.B. Jackson[5], Chris Pike[2], and Adrian Hilton[4]

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

Correspondence should be addressed to James Woodcock (`j.s.woodcock@salford.ac.uk`)

ABSTRACT

Object-based audio can be used to customize, personalize, and optimize audio reproduction depending on the specific listening scenario. To investigate and exploit the benefits of object-based audio, a framework for intelligent metadata adaptation was developed. The framework uses detailed semantic metadata that describes the audio objects, the loudspeakers, and the room. It features an extensible software tool for real-time metadata adaptation that can incorporate knowledge derived from perceptual tests and/or feedback from perceptual meters to drive adaptation and facilitate optimal rendering. One use case for the system is demonstrated through a rule-set (derived from perceptual tests with experienced mix engineers) for automatic adaptation of object levels and positions when rendering 3D content to two- and five-channel systems.

0 Introduction

Object-based audio is an approach to representing an audio scene in which individual sounds are transmitted separately along with metadata describing properties of the objects and the overall scene. This approach differs from traditional audio transmission, where the sound scene is represented by audio signals that have been rendered to drive a specific configuration of loudspeakers. There is currently an ongoing movement away from channel-based distribution towards object-based content; for example, the MPEG-H format [1] has been implemented for the Advanced Television Systems Committee (ATSC) 3.0 broadcasting standard [2].

Broadcasting each sound source or collection of sources as objects provides the opportunity to manip-

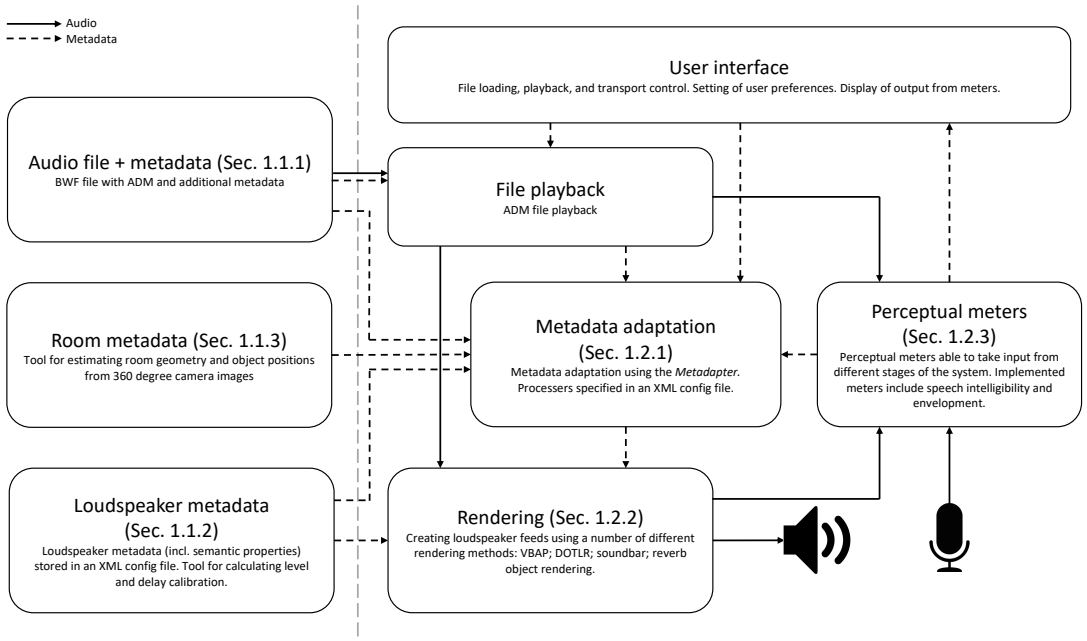

Fig. 1: Intelligent metadata adaptation framework

ulate the objects at the reproduction end of the transmission chain, meaning that the rendering of the sound scene can be customized, personalized, and optimized for the specific listening scenario (i.e., reproduction system, environment, and listener-centric factors such as preference, mood, and accessibility needs). This delay in loudspeaker feed rendering also facilitates listener control, enabling personalization such as allowing the user to select the language for speech content or changing the foreground-to-background ratio [2].

Current object-based systems generally utilize metadata relating to the intended position and level of each object. Rendering techniques such as vector base amplitude panning (VBAP) [3], ambisonics, and wave field synthesis are then used to generate loudspeaker feeds for the specified reproduction loudspeaker layout. As the rendered loudspeaker feeds are specific to the reproduction loudspeaker layout, object-based audio transmission is often said to be "format agnostic" [4]. However, it is very rarely possible to produce an identical listening experience on two different systems; for example, objects in a 3D mix that are behind or above the listener will not be positioned correctly in a two-channel stereo reproduction.

If the exact listening experience cannot be maintained, one approach to object-based audio rendering is to attempt to maximize the quality of listening experience by maintaining the producer intent as accurately as possible within the constraints of the reproduction scenario. It has recently been shown that when producers mix object-based content to different loudspeaker layouts, the mix processes that they apply depend on the semantic category of the object (i.e., whether the object is dialogue, foreground effect, diffuse background, or music) [5, 6]. Therefore, maintaining producer intent across fundamentally different loudspeaker layouts (i.e., 3D surround, 2D surround, and stereo) is likely to require complex processing. For example, how an object should be rendered might depend on the type of object or its narrative importance to the overall scene.

0.1 Aims of the intelligent metadata adaptation framework

To address the issue of optimal rendering of object-based content to different loudspeaker setups, a framework for intelligent metadata adaptation was developed. The overall aim of this work was to develop an intelligent system for customizing, personalizing, and perceptually optimizing the rendering of an object-based

audio stream to maintain the quality of listening experience across alternative reproduction setups. *Intelligent* means that the system replicates production decisions that an experienced human would make. *Customization* relates to adapting the reproduction to *any available devices*, whilst *personalization* means that users can make choices and the rendering adapts to their preferences. *Perceptually optimized* means that rendering decisions are based on perceptual models or experiment results. The aim of *maintaining the quality of listening experience* relates to creating listening experiences that are enjoyable and reflect the original producer intent.

The framework was developed from the baseline end-to-end object-based audio system presented by Coleman et al. [7], and extends current state-of-the art object-based audio rendering systems by utilizing advanced metadata about the audio objects, the reproduction space, and the loudspeakers. It also features a range of rendering techniques beyond VBAP in order to create the greatest opportunity for producing the appropriate loudspeaker feeds. Finally, perceptual meters are employed to provide relevant feedback for optimization.

A Python software package—the *Metadapter*—was developed to facilitate realtime rule based metadata adaption of an object-based audio stream. Generally the rules implemented were determined from the results of perceptual tests. For example, re-rendering strategies were implemented based on expert knowledge from mix engineers and knowledge about the perception of aspects such as envelopment, speech intelligibility, height localization and reverberation.

0.2 Paper outline

In this paper, the main components of the intelligent metadata adaptation framework introduced above are outlined. An overall description of the framework is given in Section 1, and the individual components are introduced. In order to illustrate the benefits of the system, an example use case for the system is presented in Section 2. Other potential use cases for the system and an outlook for future work is given in Section 3.

1 Description of the framework

A diagram of the intelligent metadata adaptation system is shown in Figure 1. Processes on the left hand side of the diagram are carried out offline; for example, creating audio files with metadata, generating room metadata, or determining loudspeaker metadata. Processes on the right hand side of the diagram happen online; for example, the real-time rendering and metadata adaptation stage, the user interface for control and visualization, and the perceptual metering.

1.1 Advanced metadata

As discussed in Section 0.1, advanced metadata underpins the intelligent metadata adaptation framework detailed in this paper. The framework is facilitated by detailed knowledge about the audio objects, the reproduction system, and reproduction space. This section describes the metadata used by the framework.

1.1.1 Audio object metadata

The metadata used in the system presented by Coleman et al. [7] is loosely based on the Audio Definition Model (ADM) format [8]. The ADM is an open metadata standard that specifies the description of channel-, object-, and scene-based audio. The standard specifies an Extensible Markup Language (XML) data structure that can be embedded in a Broadcast Wave Format (BWF) file. The metadata parameters available for audio objects are: *position* (in polar/spherical or cartesian formats), *width*, *depth*, *height*, *gain*, *diffuseness*, *channelLock*, *objectDivergence*, *jumpPosition*, *zoneExclusion*, *screenRef*, and *importance*.

The framework described in this paper extends the description provided by ADM by including a number of additional object- and scene-level metadata fields. The metadata added is dependent on the particular use case for the framework. An example of additional metadata related to individual objects is *objectCategory*, which describes the perceptual object categories identified by Woodcock et al. [5]. An example of additional scene level metadata is a target envelopment level for a scene. The framework is extensible; additional metadata fields that are not understood by the renderer are ignored, but they can be used to drive adaptation (see Section 1.2).

1.1.2 Loudspeaker metadata

Object-based audio systems necessarily have knowledge of the loudspeakers that are available; the renderer uses that information to generate appropriate loudspeaker feeds. In the system presented by Coleman

et al. [7], simple metadata (such as loudspeaker positions, delays, and gains) were stored in an XML configuration file. In the intelligent metadata adaptation framework proposed in this paper, advanced semantic metadata fields are added to the configuration file. Fields include *name*, *quality*, and *type*. This semantic metadata is used in metadata adaptation; for example, the *quality* field could be used to enable the routing of certain objects to loudspeakers of a specific quality in media device orchestration (MDO) reproduction [9] (see Table 1).

1.1.3 Room metadata

In certain reproduction systems, it is beneficial to know about the properties of the reproduction room. For example, knowledge of the positions and material properties of the walls could be used to adapt reproduction for an estimated reverberation time. Additionally, knowledge of the objects in the room might be used to set the positions of audio objects. Visual scene analysis methods utilizing convolutional neural networks (CNN) [10] were developed to identify the location and material of surfaces in the reproduction environments, as well as the position of objects such as furniture. The room layout with object and material information is estimated using an off-the-shelf 360° camera and the computer vision algorithm proposed by Kim et al. [11], Kim and Hilton [12].

1.2 Metadata adaptation and rendering

In order to achieve the aims of the proposed system (see Section 0.1), it is necessary to adapt the rendering based on the advanced metadata described in Section 1.1. To this end the overall rendering process is conceptually partitioned into a metadata adaptation stage and the actual audio rendering. The current framework implements these aspects in two separate processes.

1.2.1 Metadata adaptation

In the proposed system, this metadata adaptation is performed in a Python software package called the *Metadapter*. A schematic diagram of the *Metadapter* is shown in Figure 2. It receives scene metadata, modifies it through a sequence of so-called *processors* (configurable Python classes) and transmits it to the audio renderer. Processors are taken from a user-extensible processor library, and both the arrangement of the processing sequence and the behavior of the individual

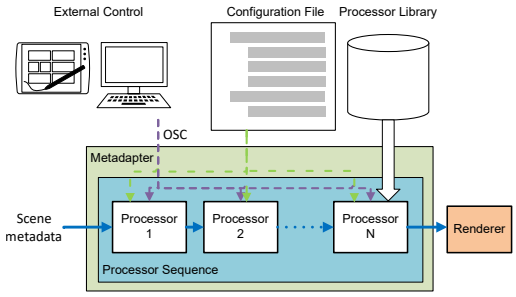

Fig. 2: *Metadapter* software framework.

processors are configured through an XML file. In addition, an OSC (Open Sound Control) interface is provided to send control messages from external user interfaces to individual processors (for example, the user interface shown in Figure 1). In this way the *Metadapter* enables extensible and interactive metadata adaptation through a concise and expressive Python syntax.

A set of processors for intelligent metadata adaptation have been implemented within the *Metadapter* framework. Most of them are based on rules determined in user tests with experienced mix engineers or perceptual tests, ensuring that the metadata processing applied is intelligent according to the definition given in Section 0.1. An overview of the implemented processors is given in Table 1. The development of one of these processors—intelligent downmixing to maintain producer intent—is described in detail in Section 2.

1.2.2 Rendering

There are many ways of generating loudspeaker feeds; however, object-based audio is often limited to VBAP, perhaps with some objects locked to particular channels. The intelligent metadata adaptation framework provides a number of different ways to render signals (detailed in Table 1), which allows selection of the most appropriate method during the metadata adaptation stage. Rendering was performed by the *Versatile Interactive Scene Renderer* (*VISR*) [20], a realtime C++ rendering framework developed within the S3A project, with the exception of the soundbar rendering, which was implemented in standalone software.

1.2.3 Metering

One of the most important advantages of object-based audio is that content can be repurposed for different reproduction systems; however, the listening experience

Table 1: Online features of the intelligent metadata adaptation framework

	Feature	Description
Processors	Add advanced metadata	Parses the extra advanced metadata chunk from the BWF files (described in Section 1.1.1).
	Azimuth remapping and level adjustment	Intelligent object-based downmixing of immersive 3D content to 2D surround or stereo (described in detail in Section 2).
	Virtual height	Utilizes psychoacoustic knowledge to create virtual height sources.
	Envelopment personalisation	Adjusts object azimuth, elevation, level, spread (i.e., the relative space between individual objects), and equalization [13] to achieve a target amount of envelopment set by the user.
	Envelopment optimisation	Attempts to match the metered envelopment to the target envelopment (from the advanced metadata) by modifying the envelopment parameters detailed above.
	Media device orchestration	Media device orchestration (MDO) is the concept of utilizing any available device for optimum reproduction of a media experience [9]. This processor implements rules to route objects to these additional devices.
	Room compensation	Modifies reverberation metadata according to a description of the acoustics in the reproduction space, to reproduce reverberant target sounds more faithfully [14].
Rendering methods	VBAP	Implementation of vector base amplitude panning, as described in Pulkki [3].
	DOTLR	Direct object-to-loudspeaker routing allows individual objects to be routed to specified channels, and is used in the MDO processor.
	Soundbar	Beamforming and transaural rendering.
	Reverb	Object-based reverberation [15].
Perceptual meters	Intelligibility	Metering of speech intelligibility [16, 17, 18].
	Envelopment	Metering of perceived envelopment using a real-time implementation of George *et al.*'s envelopment model [19].

will be potentially very different in different reproduction scenarios, and the producer will not be able to monitor every possible reproduction prior to broadcast. In actuality, this is the case for channel-based content as well; it is not possible to tell how the end-user will have configured their system, placed their loudspeakers, etc. It is therefore beneficial to add scene-level metering to a system in order that the reproduced experience can be verified and enhanced. Object-based audio provides the opportunity for repurposing the content in a perceptually optimal manner. The intelligent metadata adaptation framework supports the inclusion of perceptual models taking input from various sources (e.g., the loudspeaker feeds directly from the renderer, or the reproduced sound field captured by one or more microphones). Two such models (speech intelligibility and envelopment), based on particularly important perceptual characteristics of reproduced sound, were implemented.

2 Use case: azimuth remapping and level adjustment processor

The previous section gave an overview of the processors, rendering methods, and meters implemented in the framework. This section provides a more detailed description of the development of one of these processors: azimuth remapping and level adjustment of objects to maintain producer intent. The description of this use case is intended to illustrate some of the benefits of the intelligent metadata adaptation system. Other demonstrations and use cases that were implemented are described briefly in Section 3.

As detailed in Section 1.2, one of the main features of the system proposed in this paper is the ability to

optimize rendering by adapting the object metadata. Furthermore, the rules governing this adaption may be based on results and knowledge from perceptual experiments. One use case for this approach is intelligent downmixing of immersive object-based 3D content to 2D systems based on knowledge from experienced mix engineers. This approach is particularly useful for systems where VBAP rendering is not possible without introducing virtual loudspeakers into the configuration, such as two-channel stereo.

An experiment was conducted to investigate how experienced mix engineers downmix immersive 3D audio content to lower channel count 2D systems. Woodcock et al. [6] investigated the most common processes mix engineers employ when downmixing object-based content with the aim of preserving the producer intent of an immersive 3D reference. This study found six categories of mix processes, the most commonly used of which were changes to the level and position of individual objects. The observed frequency of use of the mix processes was found to be dependent on the semantic category of the object to which they were being applied [5]; for example, changes in position were more commonly used for transient foreground sounds whereas changes in the spatial spread of objects were more commonly used for continuous background objects.

This section details an experiment that builds on the work detailed in Woodcock et al. [6] by allowing experienced mix engineers to adjust the position and level of objects in two- and five- channel object-based downmixes of immersive 3D content to attempt to preserve the producer intent of the original content. The results of this experiment are incorporated into the intelligent rendering framework as generalized rules for downmixing immersive 3D content to two-channel stereo.

2.1 Experiment methodology

This section describes the test setup, stimuli, and method used in the experiment.

2.1.1 Participants

Seventeen participants took part in the experiment. All of the participants had experience in mix engineering, ranging from postgraduate study to working in audio production professionally. The participants were paid for their time and gave their consent to participating in the experiment.

2.1.2 Test setup

The experiments were conducted in a multichannel audio booth at the University of Salford. This room contains eighteen Genelec 8030A loudspeakers; ten speakers are located in the horizontal plane, four at approximately 30° elevation, and four at approximately −30° elevation. The speakers were calibrated to produce equal L_{Aeq} at the central listening position for a pink noise input. Three different speaker layouts were used in the experiment; these are shown in Table 2.

Table 2: Speaker positions used in the experiment. *B, M,* and *U* indicate whether the speakers are in the bottom, middle, or upper layer of the layout respectively. The numbers indicate the azimuth of the speakers in degrees.

System	Speaker positions
Sixteen channel	B-135, B-045, B+045, B+135, M-180, M-135, M-090, M-045, M+000, M+045, M+090, M+135, U-135, U-045, U+045, U+135
Five Channel	M-135, M-030, M+000, M+030, M+135
Two Channel	M-030, M+030

2.1.3 Stimuli

The experiment utilized the same programme material described in Woodcock et al. [6]. The stimuli consisted of four clips (each around 20 seconds in duration) taken from professionally produced immersive 3D audio dramas created as part of the S3A project [21]. The clips were originally mixed for a 9+10+5 channel setup and were selected to demonstrate a wide range of immersive 3D audio features. The selected clips included movement of objects in different layers of the reproduction system, enveloping and diffuse background objects, and non-diegetic music and narration. Additionally, the clips were selected to include a variety of different audio object types including dialogue, sounds indicating actions and movements, transient and continuous background sounds, music, animal sounds, and prominent transient effects [5].

The object-based scenes were played from the original production sessions in Nuendo 5.0. During the experiment the stimuli were rendered in realtime (using the *VISR* implementation of VBAP [20]) to either a sixteen-, five-, or two-channel loudspeaker layout.

2.1.4 Test method

The four clips described in 2.1.3 were presented to the participants in a random order. The participants were asked to compare a sixteen-channel rendering of each clip to a rendering with fewer channels (see Table 2). The participants were able to switch between the sixteen-channel reference and the lower channel count version at any time during the experiment.

Participants were asked to adjust the azimuth and level of the dialogue, foreground, and transient background objects and the level of the music and continuous background objects in the lower channel version to preserve the producer intent of the reference clip. Participants were explicitly instructed not to try to create their ideal mix of the clip. The reproduction was adapted in realtime using the *Metadapter* framework described in Section 1.2. For each new clip, the level and azimuth of each object in the lower channel version was reset to 0 dB and 0° respectively. The participants were also provided with a master level fader and asked to match the loudness of the lower channel version with the reference when they had finished making adjustments to individual objects. The participants completed this task for all four clips for both two- and five- channel speaker layouts meaning that each participant made a total of eight mixes. Data for each slider were logged every 0.1 seconds, allowing the participants to record automations.

2.2 Results

This section presents the results of the experiment. The position adjustment data are described first followed by the level adjustment data.

2.2.1 Position adjustments

Figure 3 shows an example of the position data collected for a single object in the experiment. The bold line indicates the original object metadata as set by the producer in the reference scene. The remaining lines show the positions set by the seventeen participants in the two-channel downmix (each line represents data for a single participant). For ease of comparison, the azimuths have been normalized to the maximum possible for the given speaker layout. As shown by the bold line, the trajectory for this object was originally intended to begin at the top back right and to end at the bottom back left. It can be seen that, although there is a large amount of inter-participant spread in the timing of the position automations, all of the participants attempted to recreate the movement of the object within the limits of the two-channel system. Similar results were observed for other moving objects (figures showing these results can be found at `http://dx.doi.org/10.17866/rd.salford.6050621`).

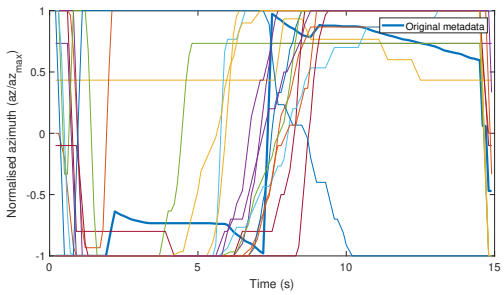

Fig. 3: Raw position data for a single object compared to original metadata.

Figure 4 shows the average azimuth of each object set by the participants over all four clips compared to the azimuth set in the original reference version of the content. As described in Section 2.1.4, position data were captured every 0.1 seconds during the experiment meaning the data presented in these figures represent both static and dynamic objects. Potentially erroneous data (e.g. movements of sliders during periods of silence) were discarded by only considering metadata when the corresponding audio object signal level was greater than -40 dBFS.

The left pane of Figure 4 shows data for the two-channel system. It can be seen that the behavior suggested in Figure 3 can also be observed in the data after averaging over the participant group; generally the participants have preserved the movement of objects behind the listener by mirroring the position in the two-channel stereo system. A candidate metadata adaptation rule to simulate this behavior is indicated by the solid red line. The Pearson's correlation between the original azimuth data and the response data collected in the experiment is 0.60 ($p < 0.001$); after applying the proposed metadata adaptation rule to the original metadata, the correlation increases to 0.77 ($p < 0.001$).

The right pane of Figure 4 shows data for the five-channel system. A linear trend between the the reference data and the response data between $\pm 135^o$ can be

Table 3: Mean level adjustments for categories of objects adjusted significantly more than 0 dB.

Object category	Two-channel (dB)	Five-channel (dB)
Dialogue	−1.6	−1.1
Actions and movement	−5.2	−3.8
Background	−2.4	−1.9
Prominent transients	−2.0	−1.9
Animal sounds	−4.6	N.S.

seen in this figure; beyond -135^o and $+135$ it can be seen that there is a tendency for objects to be panned back towards 0°. A proposed remapping function to simulate this behavior is indicated by the solid red line. The Pearson's correlation between the original azimuth data and the response data collected in the experiment is 0.80 ($p < 0.001$); applying the proposed metadata adaptation rule increases this correlation to 0.88 ($p < 0.001$).

2.2.2 Level adjustments

T-tests with Bonferroni corrections were conducted on the level adjustment data for different categories of objects for the two- and five- channel systems. These tests showed that the levels of objects in the categories *speech, action sounds, background sounds, animal sounds,* and *prominent transient sounds* were changed significantly from 0 dB by the participants after correcting for overall level adjustments to the scene. Table 3 shows the mean level adjustments in dB for the categories where significant differences were found.

2.3 Implementation of rendering rules

To illustrate how the results presented in this section can be incorporated into the intelligent metadata adaptation framework, a *Metadapter* processor (see Section 1.2) was created for remapping 3D immersive scenes to systems with only frontal speakers (i.e., two-channel stereo).

The proposed position remapping function shown in the left pane of Figure 4 was implemented using the following algorithm.

1. Mirror the object's position about the y-axis.

2. Remap to the new speaker layout using a linear interpolation as given in Equation 1:

$$\hat{\theta} = \frac{(\theta + 90)(l_{max} - l_{min})}{180} + l_{min}, \qquad (1)$$

where θ is the azimuth in the original metadata, l_{min} and l_{max} are the minimum and maximum azimuth of the speaker layout respectively, and $\hat{\theta}$ is the remapped azimuth.

When semantic metadata describing the object category are available (see Section 1.1.1), the level corrections detailed in Table 3 are applied.

Figure 5 shows the predicted position (based on a velocity vector calculated from the VBAP gains) of a moving object with and without the intelligent rendering rules applied. The object shown is intended to move from the back right of the listener to the back left. The top pane of this figure shows the intended object position and the bottom pane shows the predicated position in a stereo setup using the two rendering methods. It can be seen that without the intelligent remapping the object jumps from being panned hard right to hard left; with intelligent remapping the object retains the intended movement from the right to the left of the scene.

3 Discussion and conclusions

This paper has detailed the development of a framework for intelligent metadata adaptation in object-based audio. The overall aim of the framework is to facilitate the customization, personalization, and perceptual optimization of object-based audio to maintain producer intent and hence maximize the the quality of listening experience across alternative reproduction setups. Section 2 discussed the development and implementation of one potential use case for the system—the remapping of object positions and levels based on knowledge derived from expert mix engineers. This section briefly outlines other use cases that demonstrate the aims of the system. Finally, limitations of the system and avenues for future work are discussed.

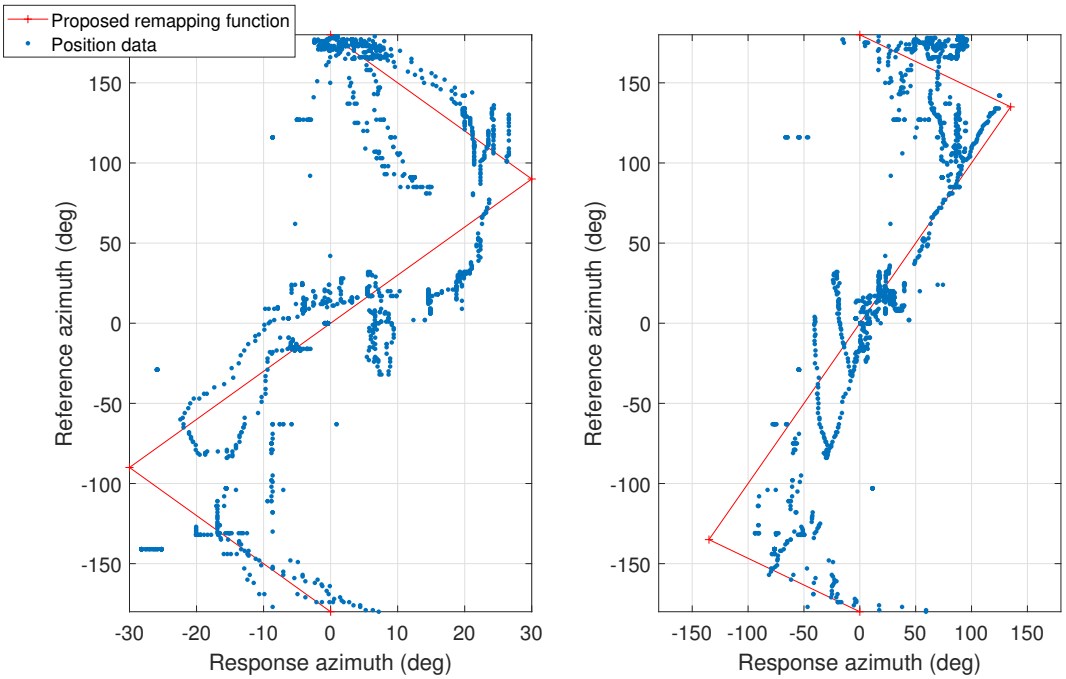

Fig. 4: Average position of each object compared to original metadata.

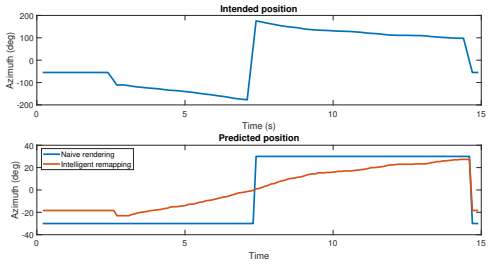

Fig. 5: Comparison of proposed rendering rules to a standard rendering.

3.1 Optimization and personalization

As detailed in Section 1.2, the system includes perceptual metering of envelopment and speech intelligibility developed using the findings of perceptual experiments. Perceptual meters can facilitate optimization and personalization by adapting the rendering to achieve a user selected or target level of the perceptual attribute. One potential use-case for this is to optimize the rendering of the scene to achieve a target level of envelopment. Envelopment was found in a previous experiment to be the most important contributer to listener preference for spatial audio systems [22].

Experiments were conducted to understand how mix engineers would adjust different parameters in an object-based mix to achieve different levels of envelopment [13]. The findings of this experiment were implemented as a set of rules to adapt the rendering to achieve different levels of envelopment. One use for these rules is to let the user control their desired level of envelopment. For example, a listener may want different levels of envelopment in different situations (e.g., high envelopment when watching a film or low envelopment for background audio whilst performing another task). In the context of the aims of the system, this demonstrates the system facilitating personalization.

Another use of this ruleset is to compensate for a loss in envelopment (e.g., when there are no rear/height speakers). A target envelopment level, set by the producer, can be stored in advanced metadata. If the predicted envelopment is different to the target envelopment, the

rendering can be modified based on the ruleset. This demonstrates the system facilitating perceptual optimization.

3.2 Customization

In the context of the system presented in this paper, customization refers to the ability to adapt the content to the devices available in the reproduction room. This was demonstrated through the use case of MDO [9]. MDO is the concept of using additional devices in the reproduction room (i.e., smart phones, tablets, Bluetooth speakers) to augment a media experience. The framework described in this paper was used, in conjunction with a combination of VBAP and DOTLR rendering (see Table 1), to develop a demonstration of the system automatically routing objects to an *ad hoc* array of consumer devices of differing quality.

3.3 Limitations and future work

The system presented in this paper provides a step towards intelligent metadata adaptation in object-based audio. There are, however, a number of limitations and avenues for future work.

Work is needed to understand the effect of the system on overall quality of listening experience. Recent research has shown that this framework in conjunction with MDO mean that an *ad hoc* array of devices can provide a statistically similar overall listening experience to a five-channel system set up and calibrated according to ITU BS.2051 [23]. Similar experiments should be conducted to assess the effect on the overall quality of listening experience of the intelligent azimuth and level adjustment rules described in Section 2 and the envelopment optimization and personalization discussed in the previous section.

Currently, the additional object- and scene-level metadata described in Section 1.1.1 are stored in an extra XML chunk in the BWF file (in addition to the ADM XML chunk) by entering the metadata into a spreadsheet and embedding it into the BWF file using a MATLAB tool. In future work, it is intended that metadata could either be added by a producer in the digital audio workstation (DAW) or determined automatically from the audio content.

The *Metadapter* processors, perceptual meters, and VISR rendering framework currently operate as separate systems. Future work should look to integrate these elements into a single deployable system though the development of plugins and VSTs that could be used in a real production environment.

Acknowledgements

This work was supported by the EPSRC Programme Grant S3A: Future Spatial Audio for an Immersive Listener Experience at Home (EP/L000539/1). The data underlying this work, along with the terms for data access, are available from `http://dx.doi.org/10.17866/rd.salford.6050621`.

References

[1] Herre, J., Hilpert, J., Kuntz, A., and Plogsties, J., "MPEG-H Audio—The New Standard for Universal Spatial/3D Audio Coding," *Journal of the Audio Engineering Society*, 62(12), pp. 821–830, 2015.

[2] Bleidt, R. L., Sen, D., Niedermeier, A., Czelhan, B., Füg, S., Disch, S., Herre, J., Hilpert, J., Neuendorf, M., Fuchs, H., Issing, J., Murtaza, A., Kuntz, A., Kratschmer, M., Küch, F., Füg, R., Schubert, B., Dick, S., Fuchs, G., Schuh, F., Burdiel, E., Peters, N., and Kim, M. Y., "Development of the MPEG-H TV Audio System for ATSC 3.0," *IEEE Transactions on Broadcasting*, 63(1), pp. 202–236, 2017, doi:10.1109/TBC.2017.2661258.

[3] Pulkki, V., "Virtual Sound Source Positioning Using Vector Base Amplitude Panning," *Journal of Audio Engineering Society*, 45(6), pp. 456–466, 1997.

[4] Shirley, B., Oldfield, R., Melchior, F., and Batke, J.-M., "Platform independent audio," *Media Production, Delivery and Interaction for Platform Independent Systems: Format-Agnostic Media*, pp. 130–165, 2013.

[5] Woodcock, J., Davies, W., Cox, T., and Melchior, F., "Categorization of Broadcast Audio Objects in Complex Auditory Scenes," *Journal of the Audio Engineering Society*, 64(6), pp. 380–394, 2016, doi:10.17743/jaes.2016.0007.

[6] Woodcock, J., Davies, W. J., Melchior, F., and Cox, T. J., "Elicitation of expert knowledge to inform object-based audio rendering to different systems," *Journal of the Audio Engineering Society*, 66(1/2), pp. 44–59, 2018.

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

[8] ITU-R, "Audio definition model," Technical Report BS.2076-0, 2015.

[9] Francombe, J., Woodcock, J., Hughes, R., Mason, R., Franck, A., Pike, C., Brookes, T., Davies, W. J., Jackson, P. J., Cox, T. J., Fazi, F. M., and Hilton, A., "Qualitative evaluation of media device orchestration for immersive spatial audio reproduction," *Journal of the Audio Engineering Society*, 2018, In Press.

[10] Eigen, D. and Fergus, R., "Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-Scale Convolutional Architecture," in *Proc. ICCV*, 2015.

[11] Kim, H., de Campos, T., and Hilton, A., "Room Layout Estimation with Object and Material Attributes Information Using a Spherical Camera," in *Proc. 3DV*, pp. 519–527, 2016.

[12] Kim, H. and Hilton, A., "3D Scene Reconstruction from Multiple Spherical Stereo Pairs," *International Journal of Computer Vision*, 104(1), pp. 94–116, 2013.

[13] Francombe, J., Brookes, T., and Mason, R., "Investigating and manipulating envelopment in object-based audio mixes," *Journal of the Audio Engineering Society*, 66(3), pp. 127–145, 2018, doi:10.17743/jaes.2018.0011.

[14] Menzies, D. and Fazi, F. M., "A Perceptual Approach to Object-Based Room Correction," in *Proc. Audio Engineering Society Convention 141, Los Angeles*, 2016.

[15] Coleman, P., Franck, A., Jackson, P. J., Hughes, R. J., Remaggi, L., and Melchior, F., "Object-based reverberation for spatial audio," *Journal of the Audio Engineering Society*, 65(1/2), pp. 66–77, 2017.

[16] Tang, Y., Cooke, M., Fazenda, B. M., and Cox, T. J., "A metric for predicting binaural speech intelligibility in stationary noise and competing speech maskers," *J. Acoust. Soc. Am.*, 140(3), pp. 1858–1870, 2016.

[17] Tang, Y., Hughes, R. J., Fazenda, B. M., and Cox, T. J., "Evaluating a distortion-weighted glimpsing metric for predicting binaural speech intelligibility in rooms," *Speech Communication*, 82(C), pp. 26–37, 2016.

[18] Tang, Y., Fazenda, B. M., and Cox, T. J., "Automatic speech-to-background ratio selection for maintaining speech intelligibility in broadcasts using an objective intelligibility metric," *Applied Sciences*, 8(1), p. 59, 2018.

[19] George, S., Zielinski, S., Rumsey, F., Jackson, P. J. B., Conetta, R., Dewhirst, M., Meares, D., and Bech, S., "Development and Validation of an Unintrusive Model for Predicting the Sensation of Envelopment Arising from Surround Sound Recordings," *Journal of the Audio Engineering Society*, 58(12), pp. 1013–1031, 2011.

[20] Franck, A. and Fazi, F. M., "VISR – A versatile open software framework for audio signal processing," in *Proc. Audio Eng. Soc. 2018 Int. Conf. Spatial Reproduction*, Tokyo, Japan, 2018.

[21] Woodcock, J., Pike, C., Melchior, F., Coleman, P., Franck, A., and Hilton, A., "Presenting the S3A object-based audio drama dataset," in *Audio Engineering Society Convention 140*, Audio Engineering Society, 2016.

[22] Francombe, J., Brookes, T., Mason, R., and Woodcock, J., "Evaluation of Spatial Audio Reproduction Methods (Part 2): Analysis of Listener Preference," *Journal of the Audio Engineering Society*, 65(3), pp. 212–225, 2017, doi: 10.17743/jaes.2016.0071.

[23] Woodcock, J., Francombe, J., Hughes, R., Mason, R., Davies, W. J., and Cox, T. J., "A quantitative evaluation of media device orchestration for immersive spatial audio reproduction," in *Audio Engineering Society Conference on Spatial Reproduction*, Audio Engineering Society, 2018.