OPEN ACCESS Freely available online

A. Franck, J. Francombe, J. Woodcock, R. Hughes, P. Coleman, D. Menzies, T. J. Cox, P. J. B. Jackson and F. M. Fazi, "A System Architecture for Semantically Informed Rendering of Object-Based Audio" *J. Audio Eng. Soc.*, vol. 67, no. 7/8, pp. 498–509, (2019 July/August.).

DOI: https://doi.org/10.17743/jaes.2019.0025

A System Architecture for Semantically Informed Rendering of Object-Based Audio

ANDREAS FRANCK,[1] *AES Member*, **JON FRANCOMBE,**[2] *AES Associate Member*,
(a.franck@soton.ac.uk)

JAMES WOODCOCK[3], **RICHARD HUGHES**[3], **PHILIP COLEMAN,**[4] *AES Member*,

DYLAN MENZIES,[1] *AES Member*, **TREVOR J. COX,**[3] *AES Member*,

PHILIP J. B. JACKSON[5], **AND FILIPPO MARIA FAZI,**[1] *AES Member*

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

Object-based audio promises format-agnostic reproduction and extensive personalization of spatial audio content. However, in practical listening scenarios, such as in consumer audio, ideal reproduction is typically not possible. To maximize the quality of listening experience, a different approach is required, for example modifications of metadata to adjust for the reproduction layout or personalization choices. In this paper we propose a novel system architecture for *semantically informed rendering (SIR)*, that combines object audio rendering with high-level processing of object metadata. In many cases, this processing uses novel, advanced metadata describing the objects to optimally adjust the audio scene to the reproduction system or listener preferences. The proposed system is evaluated with several adaptation strategies, including semantically motivated downmix to layouts with few loudspeakers, manipulation of perceptual attributes, perceptual reverberation compensation, and orchestration of mobile devices for immersive reproduction. These examples demonstrate how SIR can significantly improve the media experience and provide advanced personalization controls, for example by maintaining smooth object trajectories on systems with few loudspeakers, or providing personalized envelopment levels. An example implementation of the proposed system architecture is described and provided as an open, extensible software framework that combines object-based audio rendering and high-level processing of advanced object metadata.

0 INTRODUCTION

Object-based audio is becoming an increasingly important paradigm for producing, delivering, and reproducing (spatial) audio [1–3]. It represents audio scenes as collections of *objects*—that is, audio signals and corresponding *metadata* that describe how the object is to be reproduced. The loudspeaker or headphone signals are generated by the *renderer* using one or more algorithms. Object-based audio allows this rendering to occur later in the chain, for example, in the listener's home, than it would occur with conventional channel-based transmission. Object-based audio is, in principle, format-agnostic: the same content (i.e., audio signals and metadata) can be reproduced with different rendering methods (e.g., wave field synthesis, multichannel panning, or binaural rendering) and reproduction equipment (e.g., headphones, soundbars, or different loudspeaker layouts). In addition to this customized rendering, object-based audio offers extensive possibilities to personalize the listener experience. Potential applications include adaptation to listening modality (e.g., active or passive listening), individual preferences, or accessibility requirements (e.g., hearing impairments) [1, 4]. A number of metadata formats and standards have been proposed and standardized, including the audio definition model (ADM) [5] and MPEG-H [6].

In state of the art object-based systems, e.g., the MPEG-H reference implementation [7], the renderer typically performs low-level audio signal processing to apply a reproduction method, e.g., vector base amplitude panning

(VBAP) [8], making use of low-level core metadata such as the object level and position. This implicitly assumes that the reproduction setup, e.g., the loudspeaker layout and the reproduction room, allow the renderer to create a faithful, ideal rendering of the desired audio scene. In real listening environments, however, such ideal reproduction is often not feasible, for instance due to the number of loudspeakers, their quality, or their distribution. In such cases it would be preferable to aim at an achievable reproduction that maximizes the quality of listening experience and takes the listener's customization choices as well as the producer's intention into account. For instance, while a stereo system might not permit reproducing an object at a rear-left position, the system can ensure that it is rendered to the left because of its role in the narrative or its relation to visual objects.

This paper proposes a conceptual structure—referred to as *system architecture* in the following—for rendering object-based audio scenes. We term this architecture *semantically informed rendering (SIR)*.

As its main distinction, SIR combines conventional object audio rendering with modifications to the scene based on advanced metadata describing other aspects of the audio scene reproduction, e.g., the reproduction room, the listener including personalization choices, or perceptual characteristics of audio objects. It is based on the intelligent metadata adaptation framework introduced in [9] and extends it to a complete rendering system.

Compared to the core metadata typically used in object-based audio, this extended metadata is often more high-level, descriptive, and qualitative. For example, objects might be described by categories such as *"dialogue"* or *"background,"* allowing different metadata- or signal-level operations to be performed depending on the object category. Similarly, the quality of a loudspeaker might be rated on a scale from *"low"* to *"high."* In the proposed system, the processing of metadata is distinctly different from the low-level audio processing in conventional rendering. Metadata transformations may be derived from psychoacoustics, subjective listening experiments, or expert knowledge. For example, rules for automatic repositioning of sound objects were derived from re-mixing experiments with sound engineers [10] (see Sec. 2.1). In addition, metadata transformations can be driven by metered attributes of the sound scene, including perceptual attributes estimated by predictive models. Where such meters are driven by predictive models of perceptual attributes, they are referred to as *perceptual meters* [9].

While aspects of the metadata adaptation process have been introduced in [9], the present paper provides a comprehensive system-level view of object-based reproduction systems that enable semantically informed rendering (SIR). The main contributions are:

- *Proposing a system architecture for SIR;*
- *Definition of a metadata representation*, extending the schema proposed in [3];
- *Introduction of a processing framework for metadata adaptations*, generalizing the software introduced in [9] to foster the composition of structured, complex adaptation schemes;
- *Integration of perceptual metering* into the system architecture;
- *Description of an exemplary extensible open-source implementation* to illustrate the concepts of the system architecture, but also to enable experimentation and research into the SIR approach;
- *Discussion of the system architecture's capabilities* through a series of application examples drawn from our previous work [10–13].

The remainder of this paper is structured as follows. Sec. 1 describes the proposed system architecture, its components, their requirements, and interrelations. Throughout this section, the example implementation is used to elucidate these concepts. Applications of SIR are presented in Sec. 2, outlining their realization in the proposed architecture and the potential of semantically informed metadata adaptation. The main findings and directions for future research and development are discussed in Sec. 3.

1 SEMANTICALLY INFORMED RENDERER ARCHITECTURE

This section describes an architecture for semantically informed rendering (SIR) of object-based audio. It introduces the key components, their requirements, and interrelations. To demonstrate this architecture we present an example implementation based on the open-source VISR (Versatile Interactive Scene Renderer) framework [14].

The overall structure of the architecture is shown in Fig. 1. *Metadata and its representation* (described in Sec. 1.1) are central aspects of object-based audio and includes all forms of data (except the audio itself) that controls how audio signals are rendered. The *Metadata adaptation engine* (Sec. 1.2) is the most distinguishing part of the SIR framework. It combines metadata from all sources—including perceptual metering parameters obtained within the audio renderer—to send adapted metadata to the audio renderer. The *Object audio renderer*, described in Sec. 1.3 uses object metadata to transform a set of audio signals into channel signals for the target reproduction system, e.g., a loudspeaker setup, a soundbar, or headphones.

1.1 Metadata Representation

This section describes the metadata used in the SIR framework.

1.1.1 Metadata Model

As explained above, the use of metadata to control the reproduction of audio scenes is a main distinguishing feature of object-based audio. In this paper we use the term "metadata" to denote all descriptive data that affect the rendering, as in, for example, [2, 3, 9]. This definition is also sensible from the renderer architecture viewpoint because all these data are processed in a uniform way.

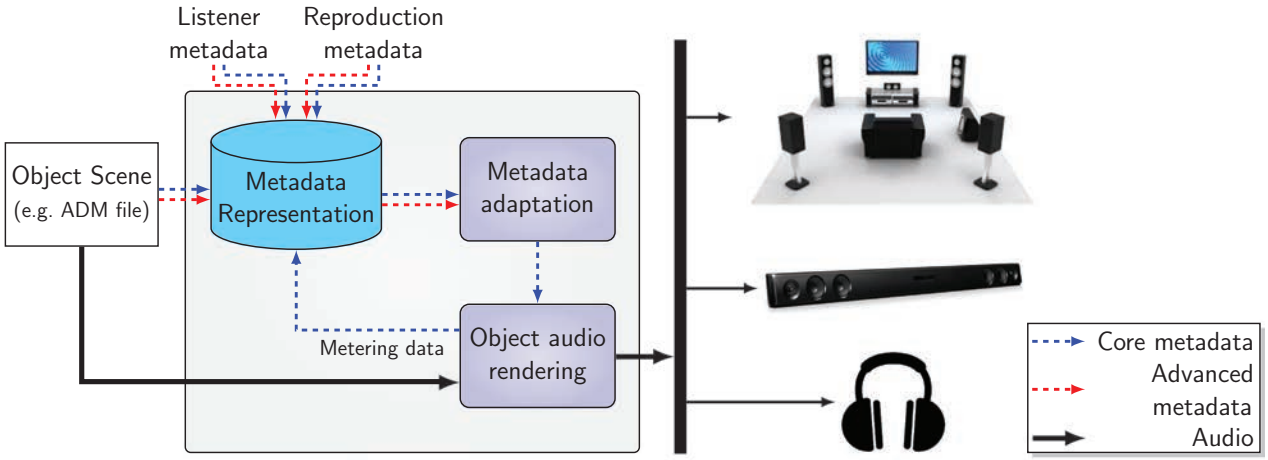

Fig. 1 Integrated semantically informed rendering (SIR) system.

Metadata for object-based audio rendering can be classified with respect to different criteria. First, parts of the metadata can be static, i.e., constant for the duration of an audio scene, or dynamic, i.e., time-varying. Examples of static data are object categories or reproduction system capabilities, while data such as object positions, narrative importance levels, or personalization choices often change dynamically.

Second, metadata can vary with respect to its level of abstraction. On the one hand, most metadata provided in current standards (i.e., ADM or MPEG-H) represent objective quantities, such as positions or sound levels, that can be directly interpreted by an object audio renderer. Here we refer to such data as *core metadata*. On the other hand, SIR makes extensive use of more abstract, often semantic, metadata not directly related to rendering parameters; for instance object categories, loudspeaker quality tags, or target values of perceptual attributes, e.g., envelopment (Sec. 2.2). These are referred to as *advanced metadata* in this paper and are often more qualitative in nature. Since they are not directly interpretable by the renderer, they are used to create and transform core metadata that can be understood by rendering algorithms. This paper focuses on the infrastructure to represent, apply, and process advanced and core metadata.

Finally, metadata can be classified by its origin:

- *Content metadata* is part of an object based scene, e.g., an ADM file or MPEG-H stream. It may describe both individual objects and scene-level properties.
- *Reproduction system metadata* describing the loudspeaker setup, the rendering system, and the reproduction room. This can include physical parameters such as loudspeaker positions, but also more qualitative data like transducer quality ratings. Parametric data about the reproduction room, such as reverberation time estimates, wall positions, and material properties can be obtained, e.g., from computer vision methods [15].
- *Listener metadata* related to the user, including listener position and listening mode as well as user-controllable inputs such as commentary language, dialogue/background balance, or the desired envelopment.
- *Metering data* generated from audio signals within the renderer, potentially including meters derived from predictive models of perceptual attributes such as intelligibility or envelopment estimates. Typically such information must be combined and processed with other metadata to affect the audio rendering.

These different classifications show that the metadata representation for SIR must be able to represent a diverse set of information from different sources.

1.1.2 Metadata Representation

Metadata are used both in the semantically informed adaptation and the audio rendering part and also forms the interface between these subsystems. Thus its representation must be suited for both parts. For the audio rendering part, only the core metadata are used, which typically correspond to a fixed data structure (for each object type) in the renderer implementation. Thus, a type-safe and efficient translation from the metadata representation into these data structures is the key requirement for this part.

For the metadata adaptation subsystem, the effective handling of both core and advanced metadata is of primary importance. This is because advanced metadata are significantly more diverse in content and structure than the core metadata. The adaptation process also requires frequent accesses and changes to that data. Therefore the representation must provide usable and convenient ways to add and manipulate complex data structures. As the manipulation of advanced metadata is typically performed at a higher level of abstraction than the audio rendering part, the data manipulation needs to support this high level of abstraction, too.

1.1.3 Example Implementation

The proposed example implementation of the SIR framework uses a metadata representation based on the JSON

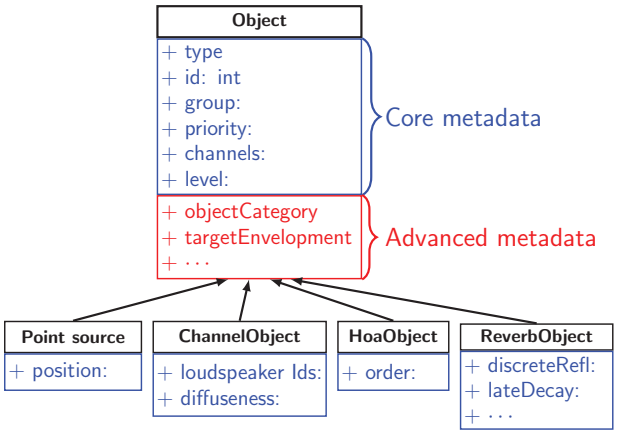

Fig. 2 Object type hierarchy used in the example implementation, extended from [3].

(JavaScript Object Notation) [16] data interchange format as described in [3]. For example,

```
{"sceneattr1": <val>,...,
"objects":[ {<obj1>},..., <objn>]}
```

is an example of a scene description. The JSON format is text-based and human-readable and describes different aspects of the scene. Here, it contains both scene-level data (e.g., the "`sceneattr1`" attribute) and object metadata in the "`objects`" array. The string

```
{"type":"point", "id":5, "chan-
  nels":"2", "priority":"0", "level":
  "0.3", "position":      {"x": "3.0",
  "y":"-0.5", "z": 0.25"},    "object-
  Category": "dialogue"}
```

is an example of an object description. Core object metadata, such as `"position,"` can be freely combined with advanced data, e.g., the `"objectCategory"` attribute. The JSON representation can be conveniently used both in the audio renderer and in the metadata adaptation subsystem.

Compared to existing metadata standards (such as ADM or MPEG-H), the proposed object format introduces two interrelated features. First, while the established implementations typically provide a single object type with multiple options, the proposed format provides a hierarchy of object types that enables each part of the scene to be represented by a matching type. A subset of this type hierarchy is shown in Fig. 2. The benefits of this approach are a more expressive syntax and a set of object parameters tailored to each type. Second, while ADM and MPEG-H support different audio representations (channel beds, "direct speaker," and higher-order ambisonics (HOA) audio) in addition to objects, the proposed description handles these representations as different object types. While this change might seem purely technical, uniform handling offers tangible advantages when combined with semantically informed metadata adaptation. For instance, it is possible to have adaptation rules to conditionally convert point sources into "direct speaker" objects, or to matrix a set of audio objects into a channel bed.

1.2 Metadata Adaptation Engine

The purpose of the metadata adaptation engine is to combine and transform all forms of metadata, including semantic information, into a set of core metadata that can be interpreted by the object audio renderer. Such metadata adaptation has specific requirements, which necessitate a separate implementation from the audio rendering subsystem. First, the level of abstraction is significantly higher. While audio rendering mainly comprises low-level signal processing operations, the metadata adaptation stages focus on transformations of object- and scene-level attributes. This requires the use of higher-level programming languages as well as access and manipulation mechanisms for metadata attributes. For example, the audio rendering typically operates with a set of fixed, static parameters; however the data at the adaptation stage are significantly more volatile, requiring language support for dynamic data types that permit the dynamic addition and manipulation of new metadata fields. Second, metadata adaptation for a complete SIR renderer often requires a number of separate adaptations covering different aspects of the reproduction. For this reason, an adaptation engine must be: (i) extensible, i.e., support the addition of new adaptation steps; and (ii) allow for the composition of multiple adaptation steps into more complex schemes.

Third, the engine should allow adaptation rules to be reused in new contexts. This also implies that the system is configurable. Finally, it needs to support the incorporation of metadata from multiple sources—such as user interactivity controls or perceptual metering data—in real-time.

1.2.1 Example Implementation: The Metadapter

In the proposed example implementation, the metadata adaptation is performed within a software package termed *Metadapter*, introduced in [9]. It is implemented in Python, resulting in readable, concise code for metadata adaptation rules. In particular, it allows direct manipulation of the metadata representation (Sec. 1.1) because its JSON data format can be directly transformed to and from dynamic Python data structures (using the data type `dict`). This makes metadata manipulation, including the dynamic addition and removal of data, easy and expressive. Control metadata can be sent to the Metadapter using Open Sound Control (OSC) messages, enabling rapid prototyping of user interfaces using tools such as MaxMSP or TouchOSC.

The structure of the Metadapter software package is shown in Fig. 3. The main parts are an extensible library of *processors*, a *processing graph* defined from a *processing configuration*, and the *adaptation engine*. *Processors* are Python classes that implement a specific metadata adaptation task. They provide a `process()` method that receives the scene metadata representation and optionally other user or reproduction system metadata, and manipulates them.

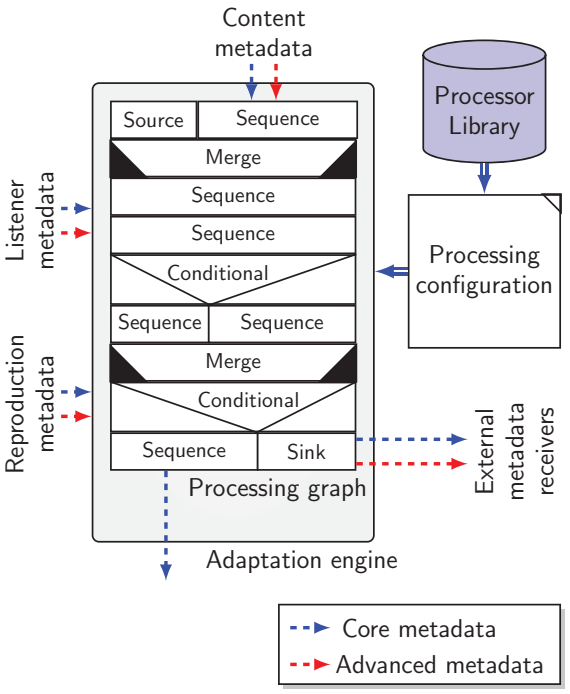

Fig. 3 Metadapter system diagram.

Most processors can be configured, allowing for effective code reuse. For more complex adaptations, processors are arranged into a *processing graph*. Its structure, together with the configuration of the contained processors, is defined by the *processing configuration*. While processors are often arranged sequentially, this graph also allows for more complex control structures, as exemplified in Fig. 3. This is enabled by a set of different processor types:

- **Sequence processors** implement the sequential processing as described above;
- **Sources** create new metadata;
- **Sinks** remove metadata elements, possibly sending them to external receivers;
- **Conditionals** (branches) select subsets of the metadata and apply different processing to them;
- **Merge** processors combine different metadata elements.

These constructs can be used to separate different aspects; for instance, selecting a subset of objects and applying a change to these objects. This enables more generic processors and improves code reuse.

The *adaptation engine* executes a processor graph, receiving metadata from multiple sources and sending the resulting core metadata on to the object audio renderer. In summary, the Metadapter enables the composition of complex metadata adaptation workflows from reusable, configurable processing elements implemented in an expressive, high-level programming language.

1.3 Audio Object Rendering

The SIR concept introduces a number of additional requirements compared to conventional object-based rendering. On the one hand, as SIR often involves multiple rendering choices for an object type, the renderer must provide means to provide multiple reproduction methods, allow for the addition of new algorithms, and enable dynamic routing to these processing facilities. On the other hand, it must allow for easy integration with the other components of the SIR architecture, e.g., to package them into manageable rendering systems.

Object audio rendering typically comprises signal operations on multichannel signals and operates under real-time constraints. Due to this lower abstraction level and the higher efficiency requirements, compared to metadata adapation, these operations are typically implemented in statically typed, compiled languages as C or C++.

1.3.1 Example Implementation: The VISR Framework

The object audio renderer of the example implementation is implemented using the VISR framework [14], an open-source software framework for audio rendering. While application-independent, its features—especially its multichannel audio architecture and its support for complex parameter communication—make it suitable for object-based rendering [3].

Fig. 4 shows an exemplary object audio renderer configuration in the VISR framework. The parts of the renderer are implemented as components, which use interconnected ports to exchange audio and parameter data. Thus, multiple reproduction methods can be implemented and added flexibly. When used as a real-time application, different forms of metadata describing the scene, listener position, and the reproduction setup, are received, decoded, and routed within the renderer. To this end the VISR software provides a library for parsing and serializing the metadata representation described in Sec. 1.1.3 as C++ or Python data types. Different rendering methods—such as VBAP, direct object-to-loudspeaker rendering, or object-based reverberation—are implemented within the object renderer. It is noted that some reproduction methods share audio processing resources, such as gain matrices for the direct and decorrelated sound. The choice of the rendering method and the routing of audio signals are controlled by metadata. Additional audio adjustments such as level changes or equalization (EQ) can be applied both to object signals and the output channel signals. These adjustments are controlled by scene and reproduction system metadata, respectively. Perceptual meters perform signal processing on audio signals to estimate perceptual attributes of the rendered scene and output them as metadata. While in the configuration of Fig. 4 these attributes are derived from the output channel signals, perceptual meters could also access other audio or parameter streams within the object renderer. In summary, a modular, component-based rendering framework is an advantageous choice for an SIR system because it is easy to combine different rendering methods and to route the complex metadata streams.

The component-based structure of the VISR framework enables an optional integration of the metadata

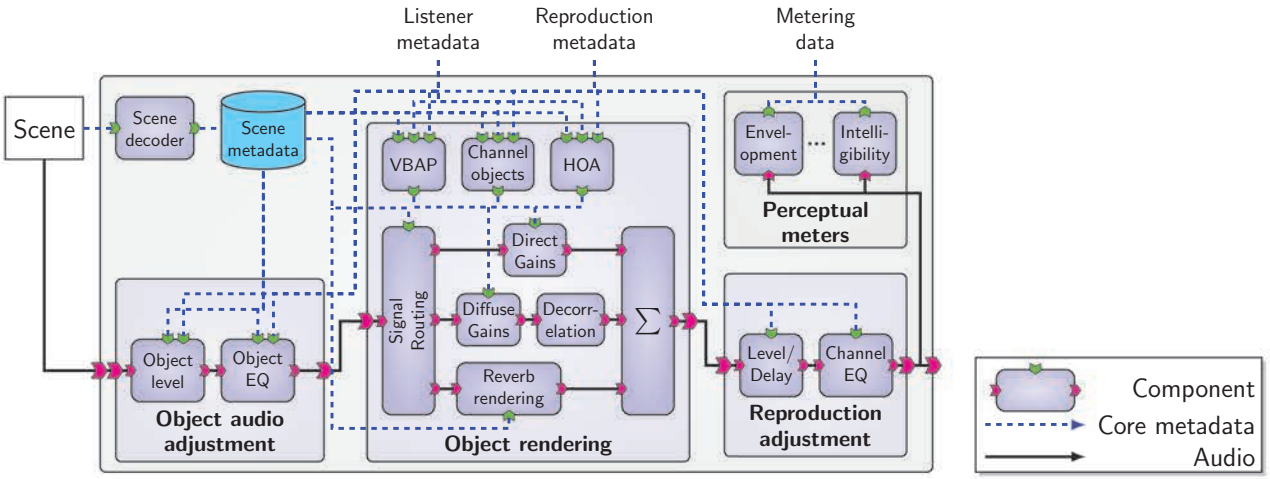

Fig. 4 Object audio renderer configuration based on the VISR framework.

adaptation into the object rendering system. In this case, the Metadapter is implemented as a component within the framework and can be configured to run different metadata adaptation schemes. While functionally equivalent to the system architecture described here, this integrated version shows how the SIR approach could evolve into self-contained end-user reproduction devices.

1.4 Summary

The system architecture proposed in this section enables flexible semantically informed rendering of object-based audio content. Built upon an extensive metadata representation consisting of core and advanced metadata, the rendering task is subdivided into a metadata adaptation stage and a lower-level object audio rendering stage. An example implementation is introduced to further exemplify this system architecture, its components, and their interrelations.

2 APPLICATION EXAMPLES

This section describes several use cases for semantically informed rendering (SIR), their implementation in the proposed architecture, and the benefits of this approach over conventional object-based rendering.

2.1 Intelligent Downmixing

One use case for the SIR approach is metadata based downmixing of immersive object-based 3D content to 2D systems. Downmixing based on metadata adaptation is particularly useful for systems where VBAP rendering is not possible without introducing virtual loudspeakers into the configuration, such as two-channel stereo. A benefit to this approach over traditional matrix downmixing methods is that it allows downmixing rules to be specified for different categories of audio objects—it has recently been shown that mix engineers apply different processes to different categories of audio objects when downmixing object-based audio to different systems [10].

To determine a set of metadata based downmixing rules, an experiment was conducted in which a group of experienced mix engineers adjusted the azimuth and level of objects in two- and five-channel renderings of content originally mixed for a 3D system. The task of the mix engineers was to produce downmixes for the two- and five-channel systems that preserved the producer intent of the original 3D reference version. Based on the outcomes of this experiment, metadata adaptation rules were derived to replicate what a professional mix engineer would do to different categories of objects when downmixing 3D content to five- and two-channel systems. Full details of this experiment are provided in [9].

The resulting metadata adaptation processor uses semantic metadata describing the category of the audio object (based on the perpetual categories detailed in [17]) and metadata describing the maximum and minimum azimuth of the target loudspeaker layout. Based on this information, the adaptation rules operate on the core metadata fields *position* and *level*.

The remapping of object positions is performed on all objects that are not continuous background sound or non-diegetic music. The remapping is performed by mirroring the object's position about the *y*-axis then applying a linear interpolation to remap the operating range of the loudspeaker layout (i.e., for two-channel stereo this is $\pm 30^\circ$). Level adjustments to each object are made based on the semantic category of the objects (see [9] for further details).

Fig. 5 shows the intended and predicted position (from the velocity vector) of a single object rendered to a two-channel layout with speakers at $\pm 30^\circ$ using standard VBAP and the SIR rules described in this section. As there is no valid VBAP solution for two-channel stereo outside the loudspeaker span, a virtual loudspeaker was included in the configuration at -180°. The energy of the virtual loudspeaker can be handled in two ways—it can be either discarded [18] or re-routed in equal proportion to the front speakers [6]. For standard VBAP, when energy in the virtual speaker is discarded the object loses any movement outside of the stereo field and jumps abruptly between speakers when the object passes behind the listener. Re-routing the energy from the virtual speaker avoids this abrupt jump, but the object still has a tendency to cling to the left or

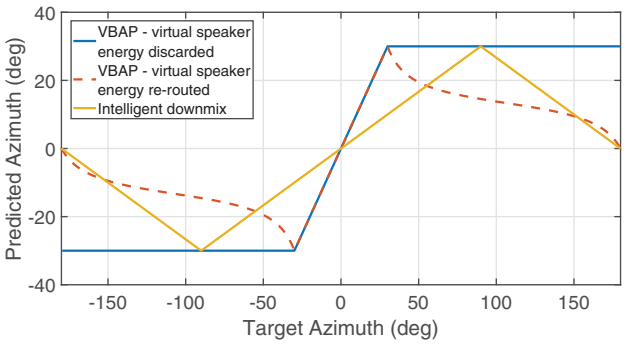

Fig. 5 Comparison of proposed intelligent downmixing rules to a standard VBAP rendering (reproduced from [9]).

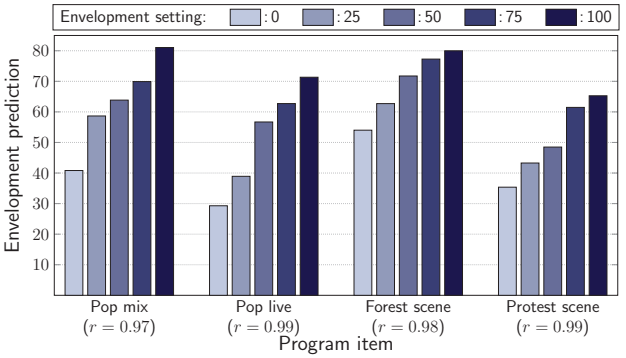

Fig. 6 Envelopment predictions for object-based mixes with personalized envelopment levels (reproduced from [11]). The r values give Pearson's correlation coefficient and are all significant at $p < 0.01$.

right speakers. By remapping the metadata to the working area of the speaker configuration according to the metadata adaptation rules, the object maintains a smooth trajectory that is closer to the original creative intent.

The use case presented in this section illustrates how the SIR framework could be used to integrate expert knowledge into metadata based downmixing. Perceptual validation is needed to understand the effect of the system on overall quality of listening experience compared to standard downmixing methods. The system could be used in a similar way to implement other common mixing processing such as EQ and reverberation.

2.2 Perceptual Attribute Manipulation

In the previous section, it was shown that semantically informed rendering of object-based audio can be used to optimize downmixing to different loudspeaker layouts, maximizing high-level perceptual attributes such as quality of listener experience and producer intent. It is also possible to adapt the rendering of object-based audio to manipulate other perceptual attributes. Envelopment has previously been shown to be a particularly important perceptual attribute [19]. Francombe et al. [11] performed an experiment to determine the relationship between parameters that can be varied in the VISR renderer (levels, positions, and equalization of certain categories of objects) and the perceived envelopment of the resultant reproduction. These relationships were coded as a ruleset in a Metadapter processor that manipulates the envelopment.

2.2.1 Envelopment Personalization

One such manipulation is the personalization of envelopment, i.e., letting a user set the level of envelopment as desired. To facilitate this control, advanced metadata describing the categories of the audio objects (as in Sec. 2.1) were added on top of the core metadata and stored within the scene. The metadata processor received a control value (user envelopment level, from 0 to 100), and adjusted the level, position, and equalization of each object in a manner determined by their object categories (see [11] for more detail). An implementation of George et al.'s model [20], which predicts perceived envelopment from a set of loudspeaker feeds, was used to show that the envelopment personalization control produced monotonically varying envelopment in a validation stimulus set (see Fig. 6).

2.2.2 Envelopment Optimization

To enable optimization, perceptual parameters could be determined (e.g., by making measurements with perceptual meters) and added to metadata at the production stage. The same parameters could be calculated at the reproduction stage and compared with the target values in metadata to assess the performance of the system with regard to those attributes. To demonstrate this workflow, a scene-level *target envelopment* field was added to the metadata. A real-time implementation of George et al.'s model [20] was configured to periodically send predictions to a Metadapter processor, which also reads the target envelopment value. The target and predicted values were compared, and the envelopment ruleset was used to modify the reproduction in an attempt to match the two.

2.3 Media Device Orchestration

In the previous sections it was shown how, in the context of the system architecture, use of advanced metadata allows the adaptation of perceptual attributes to improve the listening experience. The same system can equally be used to enable less conventional forms of spatial audio reproduction.

As explained in the introduction, most spatial audio methods pose significant challenges in creating immersive experiences in practical domestic listening environments. This is due to the quality, limited number, and restricted positioning of loudspeakers, as well as the optimal listening experience being limited to a narrow "sweet spot." As an alternative, the use of an ad hoc array of portable devices to deliver or augment a media experience—termed media device orchestration (MDO)—has recently been proposed [12]. MDO utilizes sound-producing devices that are likely to be present in domestic environments (such as mobile phones, laptops, tablets, etc.) alongside conventional reproduction systems (such as stereo or five-channel surround sound).

Francombe et al. [12] described an implementation using metadata adaptation and the VISR framework to allow rendering to an MDO system. In addition to the core object metadata, the framework allowed advanced metadata describing aspects such as target loudspeaker quality and object category (e.g., narrator, ambience, etc.) to be specified. For the loudspeakers, conventional metadata were supplemented with metadata indicating for example whether a loudspeaker is considered part of the main array or as an extra loudspeaker, as well as metadata describing aspects such as loudspeaker quality and function. A Metadapter processor allowed routing of objects to auxiliary devices based on position, quality, and function; the priority of the metadata fields used was adjustable, e.g., the system could be set to prioritize finding a loudspeaker of the appropriate quality.

Qualitative evaluation has shown an advanced metadata MDO system augmenting a traditional stereo setup to result in an overall positive listening experience, particularly for drama [12]. Thematic analysis identified three main categories relating to perceptual, technical, and content dependent aspects of the reproduction, with positive attributes relating to a sense of immersion and negative aspects largely technical in nature (e.g., quality of loudspeaker). Quantitative listening tests compared the quality of listening experience of MDO to conventional one-, two-, and five-channel systems, both on and off sweet spot [21]. In the sweet spot listening position MDO was shown to be comparable with two- and five-channel systems, while off sweet spot MDO was rated statistically significantly higher than all other systems, suggesting that MDO can reduce the dependence on the sweet spot. Both studies highlighted how the proposed system architecture with semantically informed metadata adaptation rules allowed more flexible rendering strategies for more optimal reproduction over less conventional setups.

2.4 Perceptual Room Compensation

Conventional channel-based room equalization can be used to reduce the overall room coloration. However coloration may vary significantly between the early and late parts and it is difficult to equalize for both parts separately, and harder still to achieve this over an extended area and without introducing strong artifacts elsewhere.

Synthetic reverberation is often used in audio production. This enables an alternative room compensation strategy, in which the components of each reverberation are modified at the point of reproduction [13]. An object-based representation of the production is then required, containing the dry source signals and separate descriptions of each reverberation used. This provides more freedom than channel-based equalization because each reverberant description can be modified separately, and in any possible way. To implement the process within the overall framework, the reverberation is described by metadata, which is processed by a room compensation processor, incorporating metadata about the local reproduction room.

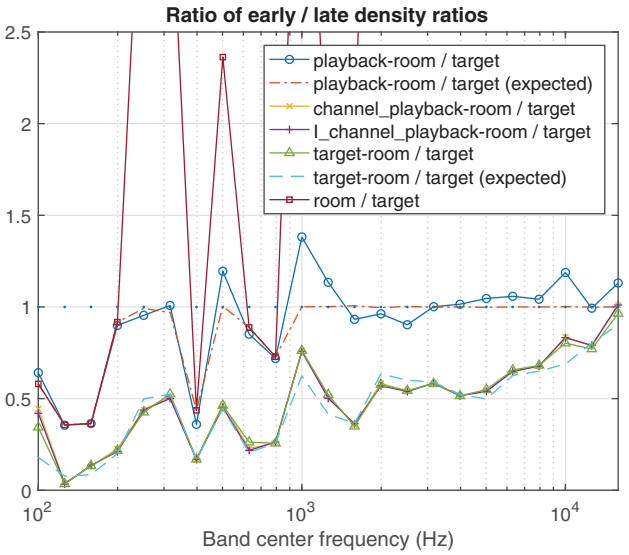

Fig. 7 Comparing the early-to-late energy ratio, or clarity, of the target against various reproductions, in 1/3 octave bands. Both target and room response are measured. *playback-room* represents the response produced by playing the processed target response into the reproduction room. Simple channel playback methods do not improve on the unmodified target played into the room.

Impulse responses are useful for representing reverberation, although they do not capture non-stationary effects that are found naturally and in some synthetic reverberators. While an impulse is relatively compact compared with an audio stream, it is also useful to represent reverberant impulse responses parametrically, as this allows the representation to focus on the most subjectively significant properties. The *Reverberant Spatial Audio Object (RSAO)* [22] is such an object-based representation, which is integrated into the VISR framework. The early response is encoded as a train of discrete reflection impulses each with direction and band equalization, and the late part is encoded with levels, attack, and decay times across frequency bands, representing diffuse sound coming from all directions.

The room compensation processor was designed to match the energies of the early and late parts of the *room-in-room* response produced if the target response were played into the reproduction room. This is performed in bands that are narrow enough both in terms of auditory perception and the smoothness of the response spectral envelopes. The RSAO metadata is naturally suited for this process. The stochastic nature of reverberance is exploited to make the processor calculation fast [13]. A paper that expands and evaluates this method is in preparation.

Fig. 7 illustrates the equalizer performance using an example case. *playback-room / target* shows the ratio of *clarities* for the object-based equalized reproduction and the target, where clarity denotes the ratio of early to late energy. The ideal ratio is 1 and is nearly achieved across much of the frequency range, but drops where the reproduction room clarity is less than the target, indicated where values of *room / target* are <1. The variation around the ideal value is because the responses are samples from a random population. More detailed, and more costly, processing can

eliminate this variation at a given point. By contrast the clarity of the channel-based and unequalized reproductions is relatively low across the spectrum, in particular at low frequencies, resulting in muddiness and reduced overall clarity.

In summary, a very efficient room equalization process has been described that leverages object-based encoding to give superior performance to conventional equalizers. The proposed system architecture provides a natural framework to house the process and integrate it within the reproduction system as a whole.

3 CONCLUSION

This paper proposes a novel system architecture for semantically informed rendering (SIR) of object-based audio, which applies extensive adaptations to metadata to improve the quality of listening experience. It proposes a separation into a metadata adaptation subsystem and an object audio rendering subsystem, describes their interrelations, and outlines the requirements for creating extensible rendering systems. This includes a discussion of the different levels of abstraction used in the two subsystems, and consequently the programming interfaces and languages that are used for each part. It also highlights the role of the metadata representation that conveys all information that affects the rendering.

Several application examples show how semantically informed metadata application schemes can be implemented in this system architecture and highlight the benefits of the SIR approach. In particular, they demonstrate how high-level metadata transformations can improve the quality of listening experience, often by incorporating expert knowledge, that would be difficult to achieve through conventional object-based audio rendering.

This conceptual change of the system architecture opens up many opportunities for object-based reproduction and personalization. First, it enables the incorporation of many developments in other fields of audio research—for example audio and scene analysis, sensing of the listeners' physiological responses, or personalization for hearing-impaired listeners—without significantly increasing the audio object renderer's complexity. Second, by separating the high-level metadata processing from the core audio rendering, it appears feasible to make significant parts of the SIR approach less dependent on the actual reproduction method. In this way, many of the envisioned metadata processing techniques can potentially be used with different reproduction techniques, e.g., loudspeaker rendering, sound bars, or over headphones. Finally, the ability to compose multiple metadata adaptation processes into larger schemes addressing different aspects of the reproduction enables the creation of complex, feature-rich systems. However, comprehensive subjective evaluation is required to show the viability and benefits of these potential future uses of semantically informed rendering (SIR).

The description of the proposed architecture is augmented by an open-source software framework that is well-suited for prototyping and creating semantically informed rendering systems. This software is provided to the research and creative communities to foster experimentation and research in this area. This includes the creation of end-to-end systems, e.g., [3], to explore, for instance, the use of metadata generated from audio-visual techniques. It also offers a perspective on how semantically informed metadata adaptation features could be incorporated in integrated reproduction systems, and, ultimately, in consumer devices that improve the quality of listening experience through object-based audio.

3.1 Future Work

Further investigation (including more perceptual testing) is needed to understand the effect of the described adaptation strategies on the perceived quality of experience. It would be of particular interest to combine multiple adaptation strategies and assess the effect. Such strategies might have complex interactions—for example, the same change might affect different perceptual attributes in different ways—and would therefore need to be carefully managed. Research and standardization effort must be undertaken to define data encodings and the technical means to integrate advanced metadata in storage and transmission formats for object-based audio, e.g., ADM containers. Finally, a better understanding of how to implement metadata adaptation strategies for alternative reproduction systems (such as soundbars or binaural headphone reproduction) is needed. This includes the design of metadata adaptation rules to improve format agnosticism.

4 ACKNOWLEDGMENT

This work was supported by the EPSRC Programme Grant S3A: Future Spatial Audio for an Immersive Listener Experience at Home (EP/L000539/1). All software and data is fully available without restriction from the DOI 10.5281/zenodo.3243995.

5 REFERENCES

[1] B. Shirley, R. Oldfield, F. Melchior, and J.-M. Batke, *Media Production, Delivery and Interaction for Platform Independent Systems: Format-Agnostic Media*, chap. "Platform Independent Audio," pp. 130–165 (Wiley, 2014).

[2] S. Spors, H. Wierstorf, A. Raake, F. Melchior, M. Frank, and F. Zotter, "Spatial Sound with Loudspeakers and Its Perception: A Review of the Current State," *Proc. IEEE*, vol. 101, no. 9, pp. 1920–1938 (2013 Sep.), doi:10.1109/JPROC.2013.2264784.

[3] P. Coleman, A. Franck, T. de Campos, J. Francombe, et al., "An Audio-Visual System for Object-Based Audio: From Recording to Listening," *IEEE Trans. Multimedia*, vol. 2, no. 8, pp. 1919–1931 (2018 Aug.), doi:10.1109/TMM.2018.2794780.

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

[5] ITU, "ITU-R BS.2076-0: Audio Definition Model," *Recommendation*, Geneva, Switzerland (2015), URL https://www.itu.int/rec/R-REC-BS.2076/.

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

[7] "ISO/MPEG 23008–3/DIS 3D Audio," International standard, ISO/IEC (2014 Jul.).

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

[9] J. Woodcock, J. Francombe, A. Franck, P. Coleman, et al., "A Framework for Intelligent Metadata Adaptation in Object-Based Audio," presented at the *AES 2018 International Conference on Spatial Reproduction: Aesthetics and Science* (2018 Aug.), conference paper P11-3.

[10] 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. 44–59 (2018 Jan./Feb.), doi:10.17743/jaes.2018.0001.

[11] 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, pp. 127–145 (2018 Mar.), doi:10.17743/jaes.2018.0011.

[12] J. Francombe, J. Woodcock, R. J. Hughes, et al., "Qualitative Evaluation of Media Device Orchestration for Immersive Spatial Audio Reproduction," *J. Audio Eng. Soc.*, vol. 66, pp. 414–429 (2018 Jun.), doi:10.17743/jaes.2018.0027.

[13] D. Menzies and F. M. Fazi, "A Perceptual Approach to Object-Based Room Correction," presented at the *141st Convention of the Audio Engineering Society* (2016 Sep.), eBrief 295.

[14] A. Franck and F. M. Fazi, "VISR – A Versatile Open Software Framework for Audio Signal Processing," presented at the *AES 2018 International Conference on Spatial Reproduction: Aesthetics and Science* (2018 Aug.), conference paper P9-2.

[15] H. Kim, R. J. Hughes, L. Remaggi, et al., "Acoustic Room Modeling Using a Spherical Camera for Reverberant Spatial Audio Objects," presented at the *142nd Convention of the Audio Engineering Society* (2018 May), convention paper 9705.

[16] IETF, "RFC 7159—The JavaScript Object Notation (JSON) Data Interchange Format," RFC (2014), URL https://tools.ietf.org/html/rfc7159.

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

[18] F. Zotter and M. Frank, "All-Round Ambisonic Panning and Decoding," *J. Audio Eng. Soc.*, vol. 60, pp. 807–820 (2012 Oct.).

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

[20] S. George, S. Zielinski, F. Rumsey, et al., "Development and Validation of an Unintrusive Model for Predicting the Sensation of Envelopment Arising from Surround Sound Recordings," *J. Audio Eng. Soc.*, vol. 58, no. 12, pp. 1013–1031 (2011 Dec.).

[21] J. Woodcock, J. Francombe, R. J. Hughes, R. Mason, W. J. Davies, and T. J. Cox, "A Quantitative Evaluation of Media Device Orchestration for Immersive Spatial Audio Reproduction," presented at the *AES 2018 International Conference on Spatial Reproduction: Aesthetics and Science* (2018 Aug.), conference paper P3-3.

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

THE AUTHOR


Andreas Franck


Jon Francombe


James Woodcock


Richard Hughes


Philip Coleman


Dylan Menzies


Philip Jackson


Trevor J. Cox


Filippo Maria Fazi

Andreas Franck received the Diploma degree in computer science and the Ph.D. degree in electrical engineering, both from the Ilmenau University of Technology, Germany. Since 2004 he has been with the Fraunhofer Institute for Digital Media Technology IDMT, Ilmenau, Germany. In 2014 he joined the Institute of Sound and Vibration Research, University of Southampton, UK as a postdoctoral research fellow. He is currently working in the EPSRC-funded project S3A: Future spatial audio for an immersive listening experience at home. His research interests include spatial and object-based audio, efficient reproduction algorithms, audio signal processing, and architecture and implementation of audio software. Dr. Franck is a member of IEEE, IEEE Signal Processing Society, and the Audio Engineering Society.

•

Jon Francombe graduated from the University of Surrey in 2010 with a First Class Honours degree in music and sound recording. He returned to Surrey in 2011 to study for a Ph.D. in perceptual audio quality evaluation as part of the Perceptually Optimized Sound Zones (POSZ) project (www.posz.org) and then worked as a research fellow on the S3A: Future Spatial Audio project (www.s3a-spatialaudio.org). Jon currently works as a senior research and development engineer in the audio research and development team at the BBC. His research focuses on audio perception, quality evaluation using quantitative and qualitative methods, and new methods of spatial audio reproduction.

•

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

•

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

•

Philip Coleman is currently a Lecturer in audio at the Institute of Sound Recording, University of Surrey, UK. Previously, he worked in the Centre for Vision, Speech and Signal Processing (University of Surrey) as a Research Fellow on the project S3A: Future spatial audio for an immersive listening experience at home. He received a Ph.D. degree in 2014 on the topic of loudspeaker array processing for personal audio (University of Surrey), as part of the perceptually optimized sound zones (POSZ) project. His research interests are broadly in the domain of engineering and perception of 3D spatial audio, including object-based audio, immersive reverberation, sound field control, loudspeaker and microphone array processing, and enabling new user experiences in spatial audio.

•

Dr. Dylan Menzies is a Senior Research Fellow in the Institute of Sound and Vibration, at the University of Southampton. Areas of interest include spatial audio synthesis and reproduction, sound synthesis for virtual environments, and musical synthesis and interfaces. He holds a Ph.D. in electronics from the University of York, an MA in mathematics from Cambridge University, and has worked as a research engineer for several companies including Sony Professional Audio.

•

Philip Jackson is Reader in machine audition at the Centre for Vision, Speech & Signal Processing (CVSSP, University of Surrey, UK) with MA in engineering (Cambridge University, UK) and Ph.D. in electronic engineering (University of Southampton, UK). His broad interests in acoustical signals have led to research contributions in human speech perception and production, auditory processing and recognition, audio-visual machine perception, blind source separation, articulatory modeling, visual speech synthesis, sound field control, and spatial audio capture, reverberation, reproduction and quality evaluation [Google Scholar: http://bit.ly/2oTRw1C]. He leads research on algorithm development for object-based spatial audio in the S3A project funded in the UK by EPSRC.

•

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

•

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