An Audio-Visual System for Object-Based Audio: From Recording to Listening

Philip Coleman, Andreas Franck, Jon Francombe, Qingju Liu, Teofilo de Campos, Richard J. Hughes, Dylan Menzies, Marcos F. Simón Gálvez, Yan Tang, James Woodcock, Philip J. B. Jackson, Frank Melchior, Chris Pike, Filippo Maria Fazi, Trevor J. Cox, and Adrian Hilton

***Abstract*—Object-based audio is an emerging representation for audio content, where content is represented in a reproduction-format-agnostic way and, thus, produced once for consumption on many different kinds of devices. This affords new opportunities for immersive, personalized, and interactive listening experiences. This paper introduces an end-to-end object-based spatial audio pipeline, from sound recording to listening. A high-level system architecture is proposed, which includes novel audio-visual interfaces to support object-based capture and listener-tracked rendering, and incorporates a proposed component for *objectification*, that is, recording content directly into an object-based form. Text-based and extensible metadata enable communication between the system components. An open architecture for object rendering is also proposed. The system's capabilities are evaluated in two parts. First, listener-tracked reproduction of metadata automatically estimated from two moving talkers is evaluated using an objective binaural localization model. Second, object-based scene capture with audio extracted using blind source separation (to remix between two talkers) and beamforming (to remix a recording of a jazz group) is evaluated with perceptually motivated objective and subjective experiments. These experiments demonstrate that the novel components of the system add capabilities beyond the state of the art. Finally, we discuss challenges and future perspectives for object-based audio workflows.**

***Index Terms*—Audio systems, audio-visual systems.**

Manuscript received August 11, 2017; revised November 20, 2017; accepted December 22, 2017. Date of publication January 17, 2018; date of current version July 17, 2018. This work was supported by the EPSRC Programme Grant S3A: Future Spatial Audio for an Immersive Listener Experience at Home (EP/L000539/1). The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Wolfgang Hürst. *(Corresponding author: Philip Coleman.)*

P. Coleman was with the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford GU2 7XH, U.K. He is now with the Institute of Sound Recording, University of Surrey, Guildford GU2 7XH, U.K. (e-mail: p.d.coleman@surrey.ac.uk).

A. Franck, D. Menzies, M. F. Simón Gálvez, and F. M. Fazi are with the Institute of Sound and Vibration Research, University of Southampton, Southampton SO17 1BJ, U.K. (e-mail: A.Franck@soton.ac.uk; d.menzies@soton.ac.uk; m.f.simon-galvez@soton.ac.uk; Filippo.Fazi@soton.ac.uk).

J. Francombe was with the Institute of Sound Recording, University of Surrey, Guildford, GU2 7XH, U.K. He is now with British Broadcasting Corp. Research and Development, Salford, M50 2LH, U.K. (e-mail: jon.francombe@bbc.co.uk).

Q. Liu, P. J. B. Jackson, and A. Hilton are with the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford GU2 7XH, U.K. (e-mail: q.liu@surrey.ac.uk; p.jackson@surrey.ac.uk; a.hilton@surrey.ac.uk).

T. de Campos is with the University of Brasilia, Gama, DF 70910-900, Brazil, and also with the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford GU2 7XH, U.K. (e-mail: t.decampos@surrey.ac.uk).

R. J. Hughes, Y. Tang, J. Woodcock, and T. J. Cox are with the Acoustics Research Centre, University of Salford, Salford M5 4WT, U.K. (e-mail: r.j.hughes@salford.ac.uk; y.tang@salford.ac.uk; j.s.woodcock@salford.ac.uk; t.j.cox@salford.ac.uk).

F. Melchior and C. Pike are with British Broadcasting Corp. Research and Development, Salford M50 2LH, U.K. (e-mail: chris.pike@bbc.co.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2018.2794780

I. Introduction

OBJECT-BASED audio representations are extremely important for future spatial audio and multimedia content consumption [1]. In the object-based audio paradigm, the content is represented as a virtual *sound scene*, which is a collection of sound-emitting *objects*. The audio for each individual object is transmitted, together with metadata describing how it should be rendered [2]. The *renderer*, part of the end user's sound reproduction equipment, interprets the object-based scene and derives the audio to be played out of each loudspeaker or headphone channel. This both ensures that the listener receives the best experience afforded by their setup, and gives new opportunities for individual listeners to personalize content. For instance, a hearing-impaired listener might adjust the balance between dialog and background sounds to improve speech intelligibility [3], or a football fan might choose to hear the match as if they are seated with their own team's fans in the stadium [4]. Furthermore, the object-based representation can allow the content itself to respond to user input based on semantic metadata [5], for instance to dynamically create a documentary of a certain length while retaining the key narrative [6].

To realize the full potential of object-based audio, system components must share common interfaces, covering the end-to-end signal pipeline from recording to listening. As there is a single, unified representation of the audio scene, which is independent of the reproduction context, the content can be said to be *format-agnostic*, i.e., in the production stage the producer is only required to create a single version of the content for all systems [7]. This, in turn, has implications for how content is commissioned, captured, produced, represented prior to rendering, and experienced by the end user. Recent standardization activity has resulted in a number of object metadata schemes. For example, MPEG-H [8] contains an object-based transmission pipeline, the audio definition model (ADM) [9] defines an extension to broadcast wave files to share and archive object-based content, and the multi-dimensional audio (MDA) [10]

is a metadata model for object-based content including a bitstream representation. These standards generally take a minimal approach to object description, and specify the object level in addition to spatial properties such as the egocentric object position and its spread, size (extent) or diffuseness. Similarly, standards are emerging for coding [8], [11] and rendering [12]; the rendering techniques are usually based on mature panning or sound field control techniques [2], [8]. The ORPHEUS project is currently investigating the object-based broadcast workflow from production to rendering, using ADM and MPEG-H [13].

One significant limitation of the above systems is that they do not fully consider the end-to-end signal pipeline. For example, an audio object usually originates by being manually created by a producer inside a production tool, rather than being directly captured from an acoustic scene. There are many situations where automated audio object creation would be beneficial, such as in live productions. Similarly, in reproduction, the state of the art workflows usually consider the rendering to be complete once the loudspeaker audio has been obtained, but make assumptions about the listeners (for instance, that they have positioned the loudspeakers correctly and are positioned in the sweet spot).

Motivated by these limitations, this paper proposes a novel end-to-end system for object-based spatial audio, integrating state of the art components to demonstrate the capability of future object-based audio systems. There are two main engineering contributions:

- *Novel system architecture:* The proposed system includes novel interfaces for: object-based capture, to allow producers to directly capture audio and/or metadata into an object-based format; visual input, to allow tracking of performers (for metadata encoding) and listeners (for sweet-spot adaptation); and perceptual meters, to monitor either the object-based scene or the rendered loudspeaker channels. Moreover, we propose an open, extensible metadata scheme for communication between components. Literature relevant to each component is reviewed. The proposed architecture will facilitate further research into the individual components as well as system-level scientific evaluations beyond the ones introduced below.
- *Open object rendering architecture:* We propose an open, flexible architecture for rendering format-agnostic content over various loudspeaker setups and over headphones. Baseline implementations of the renderer are available to the research community for evaluation and development of new object rendering approaches.[1]

The engineering contributions above lead to the following scientific contributions:

- *Evaluation of end-to-end system with visual interfaces:* We show that audio-depth performer tracking can be used to capture position metadata that allow format-agnostic rendering, and demonstrate the advantages of listener tracking with perceptually-motivated quantitative evaluation by a binaural localization model.
- *Evaluation of objectification components:* We show that objects captured, by blind source separation (BSS) for speech and by beamforming for music, facilitate personalization of channel-based recordings, where no close microphones are available. In the speech scenario, listeners increased the clarity of the target talker by adjusting the object level, while retaining acceptable audio quality compared to the channel-based reference. Objective scores for speech quality and intelligibility support the perceptual results. In the music scenario, listeners preferred mixes augmented with the object, compared to the channel-based baseline.

[1] http://cvssp.org/data/s3a/

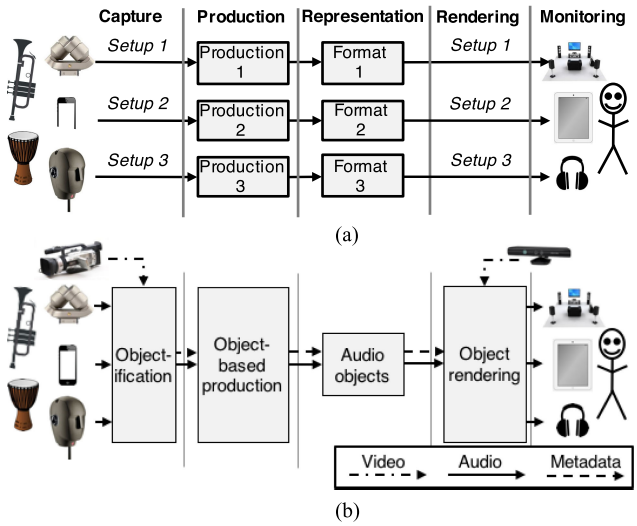


Fig. 1. Capture, production, representation, rendering and monitoring pipeline for (a) channel-based; (b) object-based audio, including the proposed *Objectification* stage and novel audio-visual interfaces for capture and rendering.

In Section II, we present an overview of the proposed end-to-end system. In Section III, components for capture are elaborated, and the production tools and perceptual meters are described in Section IV. Section V introduces the proposed object representation. Section VI describes the object rendering architecture and its application to a number of reproduction approaches. In Section VII we present several end-to-end system application examples and discuss the system's capabilities, limitations, and future opportunities. Finally, we summarize in Section VIII.

II. System Overview

In this section, we review how an object-based system differs from a traditional channel-based one, highlight the novelties in our proposed system architecture, and outline our component-based design approach.

A. Object-Based Workflow

Fig. 1 shows an overview of the end-to-end audio production chain, i.e., from acquisition to reproduction and monitoring of acoustic signals. Audio signals are first *captured* with microphones. Many different kinds of microphones can be used, such as: close microphones, where the microphone is placed near to the sound source; spatial microphone arrays, which aim to capture the overall scene including spatial information; and room

microphones, which aim to capture the ambient properties of the recording room [14], [15]. Special microphones, such as binaural microphones, ambisonic microphones, and dense microphone arrays, may also be used. In *production*, the producer brings together content captured in various formats, and, in very general terms, mixes, processes and augments the audio until the piece of content is formed. Finally, the content is encoded into a certain object-based *representation* for distribution. In the *rendering* stage, audio is sent to the loudspeakers that will reproduce the content, and the *monitoring* stage describes the content being metered or experienced by a listener.

Fig. 1 illustrates the differences between traditional channel-based audio production and the object-based paradigm. Fig. 1(a) shows the channel-based approach, illustrating that in order to achieve the best quality and maintain the producer's artistic intent, content must be recorded, produced, represented, rendered, and monitored with knowledge of the target reproduction system. For instance, stereo microphone techniques are optimally reproduced over an ideal stereo loudspeaker setup [14]. Similarly, binaural recordings can give a convincing 3D audio experience over headphones, but the effect is lost if the same two audio channels are reproduced over loudspeakers without suitable signal processing. In practice, one path is usually adopted for any given production context, with simultaneous production to multiple formats where budget constraints allow. Furthermore, there exist upmixing and downmixing techniques to translate content among different channel layouts, although these assume ideal loudspeaker setups.

On the other hand, Fig. 1(b) shows an object-based production chain. The key to such a production is the object-based representation, which is said to be format-agnostic. The proposed metadata representation, based closely on the ADM [9], is described in Section V. The representation of a sound scene by means of audio and associated metadata has implications for all other parts of the signal chain. After the content has been captured, it is immediately *objectified*, i.e., converted to a set of audio objects comprising audio and metadata (see Section III). In production (see Section IV), there are new opportunities to develop tools suitable for authoring format-agnostic content. The rendering stage (see Section VI) is critically important in an object-based pipeline, because it can exploit opportunities to optimize the listener experience for a certain system, including modification of the loudspeaker signals based on tracked listener locations, and personalization through local metadata modifications controlled by the user. Finally, the monitoring stage is exemplified in our system through perceptual meters for loudness and intelligibility. These meters are described in the context of production tools (see Section IV), but can also potentially be used within the renderer. Furthermore, our evaluations in Section VII represent a discussion on scene monitoring.

A *scene-based* production pipeline [2], for example based on Higher Order Ambisonics (HOA), would also look similar to that depicted in Fig. 1(b). For instance, the scene could initially be encoded onto HOA basis functions, manipulated and represented in this domain, and finally decoded to the available loudspeakers or binaurally rendered to headphones. The main advantage that object-based audio offers over this kind of representation is the availability of individual objects at the renderer. This retains the opportunity for personalization, giving the renderer the greatest flexibility for adapting a scene to the local reproduction system and the user's preferences.

B. Component-Based Design

The integration of the different components into an end-to-end system is one of the main contributions of this paper. We followed an approach based on component-based software engineering, e.g., [16], where the different parts of the system communicate using a set of interfaces. Specifically, these interfaces are: multichannel audio streams; JSON (JavaScript Object Notation [17]) encoded metadata; and UDP (User Datagram Protocol) network communication.

This design increases the flexibility of the proposed system. On the one hand, the individual components are interchangeable, i.e., they can be replaced by other tools and techniques implementing the same interfaces. On the other hand, the system structure can be changed easily, for instance by adding new tools or processing stages into the system.

III. Capture

The current production workflow for object-based audio requires a producer to import some audio and uses a spatialization tool to create object metadata [18], [19]. However, in future content production, it might be possible to capture media assets directly into an object-based form. This approach has been applied for experimental live sports broadcasting [20], [21], but is not yet commonplace for audio production. The proposed system offers new opportunities for object-based audio capture based on performer tracking (to obtain metadata) and the application of BSS and beamforming techniques to spatial audio capture (to acquire separated object audio). Adaptive beamforming has previously been used to capture moving talkers for reproduction by wave field synthesis [22]. Our proposed system has refined audio-visual talker tracking, extracts the scene as objects (which can be edited in production), generalizes the source separation to use BSS in addition to beamforming, and evaluates the approach for both speech and music in the context of broadcast audio. The proposed objectification stage is shown in Fig. 2. Input includes RGB-D (i.e., color + depth) video, and audio in various formats. The system blocks in the *metadata estimation* stage (see Section III-A) create metadata unsupervised or with minimal supervision, and the *supervised objectification* functionality is also provided for producers to manually add metadata to single audio channels. The *acoustic scene segmentation* stage (see Section III-B) uses audio signal processing techniques to estimate individual objects present in a recorded sound scene. Finally, the outputs of each component are collated into an object stream.

A. Metadata Estimation

The concept of automatically capturing metadata from a live recording can be transformative for object-based audio recording. The overall aim is that, by combining appropriate audio

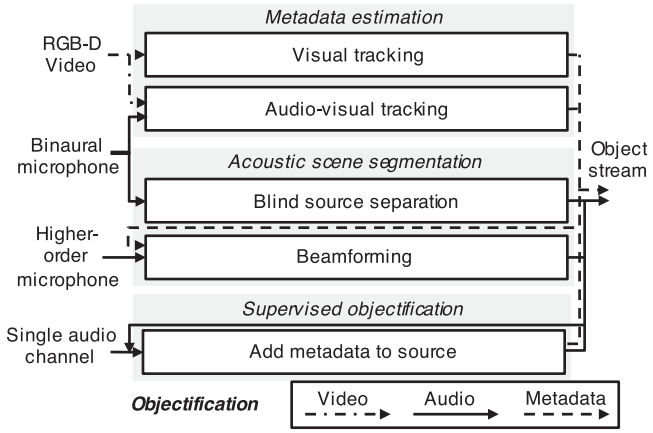
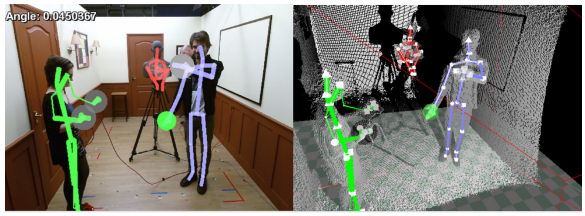


Fig. 2. System components for the *objectification* stage.

and visual inputs, objects from a live or session recording could appear in a user interface at their estimated position, and with other properties also pre-populated. The producer could then modify the scene as desired. The current version of our system is capable of tracking human performers, with applications in television, radio, and theatre recording. Similarly, our capability currently focuses on the core talker tracking technology and encoding the output as an audio object, and we have not yet developed the envisaged production tool. To date, we have developed tracking approaches using visual and audio-visual modalities.

1) Visual Tracking: Visual tracking was used both for audio caputre and reproduction. For our goal of talker objectification, the most important aspect was to estimate talkers' head locations (especially their mouths). Similarly, to apply visual tracking for listeners (see Section VI), the most important aspect is to track the head and ear locations. Therefore, our approach was to use a 3D head tracker capable of tracking multiple people, so that the same tracker could be used for both applications. A number of methods have been proposed to detect and track people in depth images [23], [24], particularly those generated using sensors based on structured light projection. We performed a set of preliminary experiments comparing state of the art implementations of 3D head tracking from depth measurements such as the method of [25], and color image (RGB) methods, such as [26]. These methods were compared against Microsoft Kinect for Windows v2.0[2] (termed *Kinect2* hereafter), a state of the art commodity RGB-D sensor that emits near infra-red pulses and estimates depth based on phase differences. We used the skeletal tracking implementated in the sensor's native software development kit; our qualitative observations indicated that this method achieves state of the art accuracy in head position estimation while being more robust than other real-time methods, as it swiftly re-initializes tracklets after occlusions. Furthermore, since it was designed to work in living rooms, the range of distances where it operates (0.5–4.5 m) is well suited to our applications. Fig. 3 shows the *skeletons*, i.e., leg, arm,

[2]URL: http://www.microsoft.com/en-us/kinectforwindows/meetkinect/features.aspx (checked 10/2016)



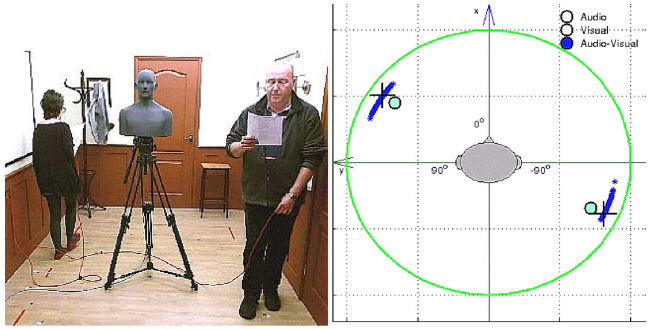
(a) (b)

Fig. 3. RGB and depth map showing skeletal tracking results of performers (or listeners) on an RGB-D frame. (a) RGB. (b) Depth map.

torso and head position estimates, detected in a sample image from Kinect2, showing (a) RGB and (b) depth channels. We integrated this tracking method into our system by streaming the skeleton positions over a network via UDP, using a structured JSON text string. For performer tracking (unsupervised metadata capture), the performer positions were converted to audio object metadata and linked to the corresponding audio (recorded using close microphones or acquired using unsupervised object separation techniques). The same JSON format is used in reproduction to inform the object-based renderer of listener positions (see Section VI).

2) Audio-Visual Tracking: Audio information can improve the robustness of visual-only tracking, and in particular can help overcome limitations in the visual data such as poor illumination and occlusions [27]. An overview of the current state of audio-visual fusion can be found in [28]. We developed a novel cross-modal person tracking algorithm combining information from a visual depth sensor and simultaneous binaural audio recordings. We acquired data using Kinect2, as above, and a Cortex Manikin MK2 binaural head and torso simulator (Cortex MK2 HATS). Fig. 4(a) shows a video frame used for audio-depth tracking.

When there are occlusions in the visual data, inconsistencies are observed in the depth head tracking results both spatially (clutters) and temporally (mis-detections). To remove clutters, we used a modified probability hypothesis density (PHD) filtering method [29] with an adaptive clutter intensity model, which takes into account measurement-driven occlusion detection as well as the depth sensor's field of view. After PHD filtering, we applied an identity (ID) association scheme [30], to ensure that the detected ID of each tracked person was consistent throughout a whole scene. Finally, to compensate mis-detections, additional information extracted from the binaural recordings was exploited. The audio-depth fusion method contains three steps. First, within *segments* (i.e., the time periods that contain trajectories without interruption), trajectory constraints for each detected target are learned via plane fitting, under the assumption that head positions from the depth tracker lie on a plane. Second, given the HATS position, the azimuth of each target relative to the HATS is calculated via depth-azimuth mapping, using 3D points projected on the plane associated with the target. Third, during each gap between segments, time-difference-of-arrival (TDOA) cues are calculated by comparing the difference between the binaural microphones, via the generalized cross correlation (GCC) [31]. Audio azimuths can be obtained from these TDOA cues based on a third-order polynomial mapping, which extends to 3D locations using a gap filling technique,



(a) (b)

Fig. 4. Illustration of audio-depth tracking for two talkers, showing the video frame and the corresponding position estimates. The estimated locations were directly encoded into the audio object metadata for unsupervised objectification. (a) Video frame. (b) Localization plot.

where the learned trajectory constraints and depth-based azimuths enclosing this gap are enforced.

The proposed PHD filter for audio-depth tracking greatly mitigates outliers in the skeletal motion. Quantitative evaluations on a dataset recorded in a typical living room showed that the average outlier rate (the number of frames containing either mis-detections or clutters) was reduced from 4.45% to 1.82% after PHD filtering. In addition, the average recall (the number of frames in which the person was successfully detected over the total number of frames containing active talkers) increased from 81.9% in the original recordings to 91.3% after applying our proposed multimodal tracking method. Fig. 4(b) shows a localization plot obtained by audio-depth tracking. In the proposed system, the position information was linked to the corresponding audio and used to populate audio object metadata (see Section V), as described in the application example in Section VII-B.

B. Acoustic Scene Segmentation

In certain scenarios it is not practical to record all individual sound sources, and instead one or more microphones in an array might be positioned to capture the overall sound scene. In order to directly capture audio objects, these signals can be processed to estimate the audio from the sources in the scene. An early investigation into audio object separation was conducted in [32], where experiments applied BSS and deconvolution to separate speech from two talkers in a conference room. In our system, we investigate BSS alongside beamforming for the application of object-based audio capture.

1) Blind Source Separation: In situations where it is impractical to place a microphone near sound sources, such as multiple talkers in a theatre production, techniques from BSS can help to estimate audio objects due to the individual talkers, facilitating remixing in post-production. Potential tools include independent component analysis [33], sparsity-based time-frequency (TF) masking [34], [35], non-negative matrix factorization [36] and deep neural networks [37]. In our system we applied a TF masking method [35]. This method exploits the interaural level differences (ILD) and interaural phase differences (IPD) in the spectral domain, while enforcing a sparsity constraint that each TF cell is dominated by at most one sound source, which is a valid approximation especially for speech signals. We used the Cortex MK2 HATS to acquire simultaneous speech from two talkers in a reverberant room, and evaluated the BSS performance. Perceptual evaluation of speech quality (PESQ [38]) scores in the range 2.2–2.6 were obtained for separated sources (over different test phrases). When the separated sources were re-mixed with different gains, as would be the case in a typical audio production, the PESQ scores increased up to 3.0, which indicates 'fair' audio quality. These experiments are discussed fully in [39]. We also applied BSS to remix a scene with two talkers, while maintaining acceptable audio quality, in a perceptual test described in Section VII-C.

2) Beamforming: Array signal processing techniques can also be used to estimate object audio signals due to a number of sources in a sound scene. In general, the spatial processing requires the source position(s) *a-priori*, which could be provided by the producer (live or during post-processing) or by using output from the tracking techniques described above. Once the source directions with respect to the array are known, the microphone array can be steered towards sounds from the target directions and suppress sound from other directions. Many excellent reviews of beamforming techniques are available [40]–[42]. We evaluated the ability of a number of classical additive beamformers to extract individual objects from a sound scene [43]. For target speech, delay and sum beamforming on the data described above gave a PESQ score of 2.35, compared to 1.99 for a single reference omnidirectional microphone. The approaches with more complex cost functions (e.g., data-based processing and spatial null creation) tended to introduce further artefacts. However, a microphone array deployed in a sound scene could still be useful to derive audio objects to send alongside channels, allowing the scene to be edited when no close microphone signals are available. We investigate this application in Section VII-C.

C. Discussion

We propose the concept of objectification, i.e., direct acquisition of acoustic signals into an object-based representation. Metadata was captured by tracking performers and converting the tracked positions to our proposed streaming metadata (see Section V). Audio corresponding to individual objects was estimated from audio mixtures using BSS and beamforming. The latter aspect of objectification is a very challenging process for real-world scenes, even using state of the art approaches. In reverberant rooms, a dereverberation front-end may be a useful addition to the objectification pipeline, as pre-processing to the approaches implemented here. The state of the art in dereverberation has recently been summarized and evaluated in relation to the REVERB challenge [44]. In addition, it would be beneficial to access a purpose-designed perceptual model; PESQ only covers speech and can only evaluate the quality of individual objects, not a whole scene. Informal listening with the BSS-estimated audio used in [39] revealed that we were able freely to respatialize these objects without exposing the interfering speech (although there were audible degradations to the target audio). Thus, we conclude that while state of the art techniques cannot provide sufficient audio quality to cap-

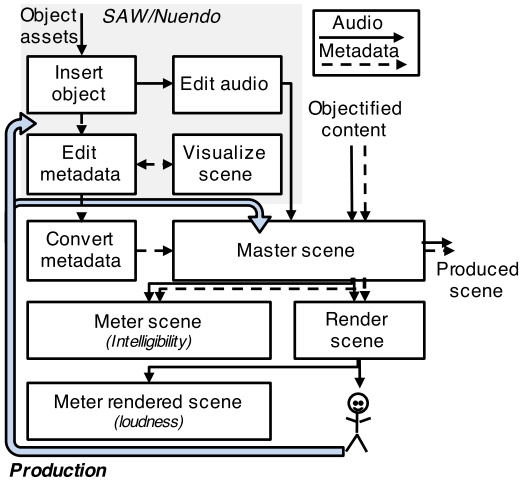


Fig. 5. System components for the production stage, showing the *baseline tools* (Spatial Audio Workstation (SAW) and Nuendo), *scene mastering*, production rendering and *perceptual scene metering* tools.

ture a purely object-based scene at this time, they do already provide sufficient interference rejection to allow remixing or respatializing. Nevertheless, BSS and beamforming can be used to estimate object signals to be transmitted alongside channel feeds and allow these scenes to be manipulated in a meaningful way without unacceptably degrading the audio quality, as the channel signals may mask any artefacts of the source separation. Current standards [8], [11] support this kind of hybrid scene. These applications are explored in Section VII-C.

IV. Production

In object-based production, the audio assets are sent directly to the renderer, while the spatialization tools operate on the metadata and not on the audio (as would be the case for a panning-based control in channel-based production). The production tool components in the system are shown in Fig. 5.

A. Baseline Tools and Mastering

Our system used a commercial solution, the Spatial Audio Workstation (SAW)[3], as the main object-based production tool. The SAW is a plugin for Nuendo[4] that includes an interface for positioning objects in 3D, and generates metadata that we subsequently converted to our JSON representation and sent over UDP. The audio channels corresponding to each object were sent to the renderer over a MADI [45] connection. Conceptually, the SAW is simultaneously used here for *supervised objectification* (see Fig. 2) and for the producer to set object positions in the metadata as part of *object-based production* [see Fig. 1(b)].

To combine the object streams from the SAW with output from the proposed objectification tools, we created a simple Max/MSP[5] program to aggregate the two metadata streams into a single stream (the audio was routed directly to the renderer). This program received both object streams and stored the values into an internal dictionary representation, before making the attributes available for editing and finally writing the full scene into metadata and sending it to the renderer. We also incorporated a simple metadata-based mastering control, whereby a different gain could be applied to an object's level field depending on the value of the corresponding priority field. This approach could also enable the end user to personalize their listening experience in a practical system.

B. Perceptual Scene Metering

The object-based audio workflow presents opportunities for perceptual metering. Our system includes interfaces for meters based on the object scene and the rendered scene. We exemplified these interfaces by implementing meters for speech intelligibility and loudness.

1) Speech Intelligibility: Speech intelligibility is naturally an important consideration for audio scenes with dialog, especially for hearing-impaired listeners. A meter based on the binaural distortion-weighted glimpse proportion metric (BiDWGP) [46] was integrated into the proposed system. The output of BiDWGP is an index falling into the range 0–1, with a larger number indicating higher intelligibility. In our prototype, this value was presented directly to the producer. The meter directly accepts the object audio and metadata as inputs, and creates an internal binaural representation for input to the model. This provides an approximation of intelligibility of the object-based scene design prior to rendering. The meter does not at this stage account for degredations due to the limitations of a particular renderer, or other local environmental noise. Nevertheless, we believe ours to be the first object-based speech intelligibility meter integrated into a 3D object based mixing workflow.

2) Scene Loudness: Loudness for multichannel audio has recently been studied [47], [48], and the International Telecommunication Union (ITU) have standardized an algorithm for predicting the perceived loudness of reproduction systems with an arbitrary number of loudspeakers at any position [49]. The loudness meters implemented in the proposed system are based on this standard and predict loudness either from (i) the loudspeaker feeds, or (ii) a binaural auralization of the loudspeaker reproduction (see Section VI-C). In (i), the model receives the rendered loudspeaker feeds and the loudspeaker positions, and uses the coefficients specified in [47]. This approach allows the producer to assess changes in loudness caused by rendering to different loudspeaker setups, and to mix the scene accordingly. In (ii), the model receives the binaural signal and predicts its loudness using a modified version of [49], in which the head-shadowing filter was bypassed [50]. The binaural input facilitates assessment of the loudness of a direct binaural render or metering of real-world setups using binaural auralization. Francombe *et al.* [48] showed similar accuracy in both of these approaches to loudness prediction (with marginally better performance statistics for the loudspeaker feed model).

V. Representation

The object-based scene representation (i.e., the object metadata) is a central part of the end-to-end system, because it links

[3]URL: http://www.iosono-sound.com (checked 10/2016)

[4]URL: http://www.steinberg.net/en/products/nuendo_range/nuendo/start.html (checked 10/2016)

[5]URL: https://cycling74.com/products/max/ (checked 12/2016)

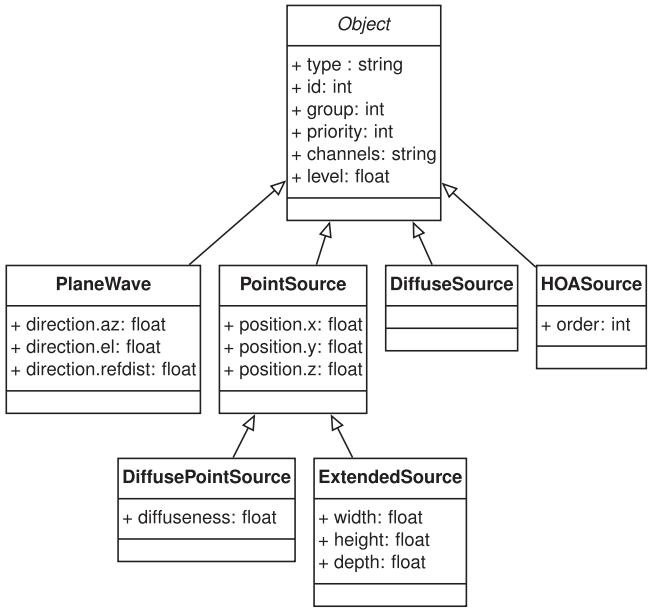


Fig. 6. Object metadata representation type hierarchy.

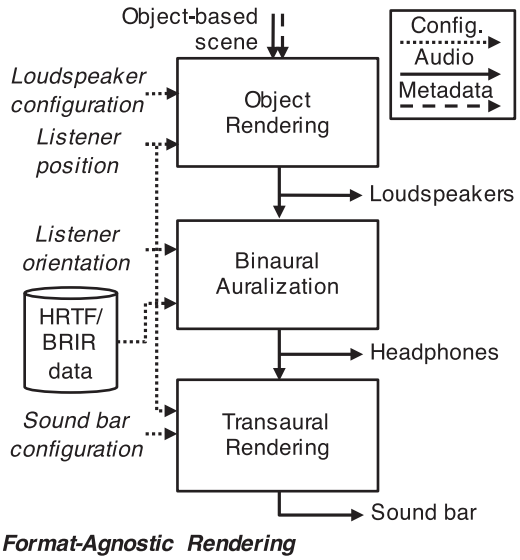


Fig. 7. Block diagram of the format-agnostic rendering.

all stages of the system. Existing object-based scene representations in audio research include the Audio Scene Description Format (ASDF) [51] and SpatDIF [52]. We chose the JSON-based representations over these formats, as they are either purely file-based (ASDF), or their network encoding based on open sound control[6] limits extensibility and the addition of semantic metadata (SpatDIF). The proposed object representation is loosely based on the ADM [9], in particular regarding object features and attributes, but there are significant differences. Firstly, ADM is a static description of a complete scene, including its temporal behavior, contained as an XML representation within a Broadcast Wave Format file. In contrast, the proposed object scheme is a streaming representation where object data is transmitted repeatedly, typically over a network connection, to convey the time-varying state of the audio scene. Secondly, while existing metadata representations as in ADM or MPEG-H feature only a single object type with several, often inactive attributes, the proposed object format features an extensible hierarchy of object types. The motivation of this type hierarchy is the ability to represent different parts of the audio scene using the best-matching object representations, to utilize the rendering resources efficiently, and to add new object types and corresponding rendering techniques as the system evolves. Recently, the reverberant spatial audio object [53] has exploited this extensibility.

Within the proposed system, we chose a metadata representation based on JSON. This text-based representation was chosen to allow for human readability, convenient metadata transformations, easy incorporation of new metadata, and extensibility. As a text-based format, the proposed representation does not at this stage put emphasis on the required bandwidth or coding efficiency, as done, for example, in transmission formats such as MPEG-H [54], MDA [10], or AC-4 bitstreams [11]. Fig. 6 shows the object type hierarchy. The base type Object contains attributes shared by all types: an object id, type label, the associated audio signal channels, grouping information, and a priority value. Currently supported object types are plane waves, point sources, and diffuse sound events. The Diffuse Point Source and Extended Source types extend the basic Point Source type by specific attributes, i.e., a controllable amount of diffuse sound or a physical source extent, respectively. A higher order Ambisonics (HOA) object type is also provided to represent sound fields, similar to the scene container in MPEG-H.

[6]URL: http://opensoundcontrol.org

VI. Rendering

The object-based renderer transforms the object audio signals and metadata, together with additional control data such as the listener's position and preferences, into loudspeaker or headphone signals for the actual reproduction configuration. Fig. 7 shows the overall signal flow of the proposed rendering system. The *Object Rendering* stage forms the core of the system. Its architecture is described in Section VI-A, and the currently implemented rendering algorithms are outlined in Section VI-B. The *Binaural Auralization* module, which can be used to present the output of a loudspeaker rendering over headphones, is described in Section VI-C. Section VI-D describes the implemented *Transaural Rendering* as an additional reproduction method to reproduce object-based binaural audio over loudspeaker arrays.

A. Software Renderer Architecture

The object renderer follows the principles of component-based software design outlined in Section II-B. To this end, it is implemented as a modular, extensible, and portable C++ framework. Within this framework, the rendering algorithms are modeled as a signal flow consisting of interconnected active elements, termed *components*. Fig. 8 shows the signal flow of the current object renderer. Components can represent either configurable, generic functionalities such as gain matrices,

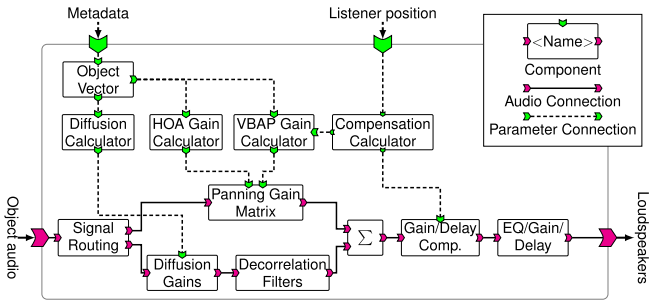


Fig. 8. Signal flow of the proposed object renderer.

delay lines, or multichannel filtering kernels, or bespoke functionality as the gain calculation of a specific panning algorithm. Components interchange both audio data and parameter data of arbitrary types, which range from complete object metadata to low-level parameters such as matrix gains or filter coefficients.

In contrast to other frameworks for multichannel audio, such as the SoundScape Renderer [55], the proposed software rendering platform focuses on the implementation of the signal processing algorithms and their interaction to form complex rendering schemes. While the former places much importance on the distribution of low-level processing tasks to multiple processor cores, the proposed software framework handles tasks such as the transmission of audio and parameter data, as well as parallelization, in runtime libraries which do not require user intervention. This enables rapid development and refinement of complex signal flows, and makes it easier to apply the same approach on different platforms. The key signal flows for object-based rendering are described below.

B. Object Rendering

The overall structure of the object-based renderer reflects the hierarchical object model described in Section V, which results in specific processing resources and signal flows for the different object types. A number of components form the infrastructure common to all object types. This includes the `Object Vector`, which represents the decoded metadata, the `Signal Routing` block that distributes the object audio signals to the processing resources, and the summation block that combines the loudspeaker signals of the different rendering approaches. The `EQ/Gain/Delay` component adjusts the output signals to the actual loudspeakers, and the optional `Compensation Calculator` with `Gain/Delay Compensation` functionality is used for adapting to listener position. In the following the core reproduction methods used for object rendering are described. Although the representation and architecture support multichannel HOA objects (which are decoded to the loudspeakers using the All-round Ambisonic Decoding (AllRAD) [56] approach), here we focus on the rendering of objects corresponding to a single audio channel.

1) Point Source and Plane Wave Rendering: Point sources and plane waves represent localized, point-like objects, at finite or infinite distances from the listener, respectively. They are among the key object types to model sound scenes. The proposed system uses Vector Base Amplitude Panning (VBAP) [57] for these object types. As standard VBAP considers only the object's direction, and not its distance, both object types are rendered identically unless the listener-adaptive panning (see Section VI-B3) is active. The implementation is partitioned into the `VBAP Gain Calculator` and the `Panning Gain Matrix` components. The former divides the spherical loudspeaker setup into a mesh of triplets and inverts a gain matrix for each triplet at startup time. The panning gains are calculated at runtime by selecting the matching loudspeaker matrix and multiplying it with the object position. The `Panning Gain Matrix` applies these gains to the object's audio signal to form a set of loudspeaker signals, ensuring good audio quality by providing smooth gain transitions in case of position changes.

2) Diffuse and Spread Object Rendering: The object types Diffuse Source and Diffuse Point Source represent either fully diffuse, omnidirectional sound events or directed sources with a adjustable fraction of diffuse energy, similar to the "spread" parameter in ADM [9]. The `Diffusion Calculator` computes a diffusion gains, which are used by the `Diffusion Gains` component to mix the object signals into a mono downmix. Finally, the `Decorrelation Filters` component applies a bank of random-phase allpass filters to create decorrelated loudspeaker feeds that are combined with the panned objects to form the final loudspeaker feeds.

3) Listener-Adaptive Panning: The rendering methods also incorporate a listener tracking functionality (described in Section III-A1 in the context of performer tracking), to fix images in absolute space. Parallax cues can then provide an improved overall sense of space as the listener moves. This was previously described for stereo panning [58], [59]. When the listener moves, the egocentric angular locations of the loudspeakers change, so the VBAP triplet inverse matrices are recalculated. In the signal flow (see Fig. 8), this is represented by the Compensation Calculator component that transmits listener position updates to the VBAP Gain Calculator block. The vectors pointing to the loudspeakers and sources are recalculated by subtraction of the listener position. The delays and gains of the final feeds are compensated for the distance of the listener to the corresponding loudspeakers by the Gain/Delay Compensation component.

C. Binaural Auralization

In the proposed system, we also implemented a binaural auralization stage, as shown in Fig. 7. The objects are first rendered to a loudspeaker setup, and the channel feeds are auralized. This provided a good spatial impression and required a sparse mesh of measured head-related transfer functions (HRTFs), compared to directly spatializing each object. Defining the system interface in this way enabled us to interchangeably use anechoic HRTF sets with virtual loudspeakers [60] or measurements of specific loudspeaker systems installed in listening rooms. Binaural signals provided signals for transaural processing, and for production meters.

D. Transaural Rendering

The proposed system architecture also allows the object-based scenes to be reproduced over a soundbar. The interface

to transaural rendering (binaural reproduction with loudspeakers, Fig. 7) is especially important because soundbar systems can feasibly be deployed in living-room type listening environments. Transaural audio has been studied using arrangements of two loudspeakers [61], leading to the development of geometries such as the stereo dipole [62] and the optimal source distribution (OSD) [63]. Recent advances in transaural reproduction have incorporated head-tracking and hence adapt the audio reproduction to the instantaneous listener position. Several works have been carried out in this area, using ensembles of loudspeaker pairs [64] and also larger loudspeaker arrays [65]. The implemented listener-adaptive transaural processing stage is fully described in [66]; cross-talk cancelation filters were created for the on-axis listening position and combined with a network of delays that steer the array output according to the instantaneous listener position (assuming that the listener is facing the center of the array). Informal listening showed the spatial impression to be convincing even when listening off-axis.

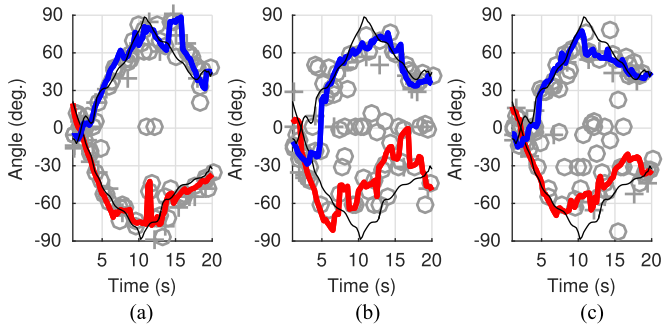


Fig. 9. Predicted localization of a two-object scene captured with automated metadata estimation and rendered over a "9 + 10 + 3" loudspeaker setup. DOA predictions were made using binaural auralizations of the 22 loudspeakers with (a) anechoic HRTFs, (b) measured BRIRs in a listening room in a position outside the sweet spot, (c) as (b) but with the listener tracking compensation active. Candidate DOAs (first ° and second + peaks detected from the DOA histogram) are shown, alongside the estimated talker trajectories (thick lines) and the ground-truth metadata positions (thin black lines).

VII. Discussion: Applications and Opportunities

In this section, three main applications of the proposed system are evaluated: scene composition with clean object audio and manual metadata authoring, scene capture with automated metadata, and scene capture utilizing audio segmentation. Then, we discuss the overall capabilities and limitations of the system, and highlight key areas for future work.

A. Manual Scene Recording and Authoring

An early version of the proposed system, prior to the integration of the objectification component, was utilized in the production of three object-based audio drama scenes [67]. Dialog for the scenes was recorded in a semi-anechoic environment; each actor had a separate microphone and they were asked not to overlap their lines. Additional objects utilized in the final scenes were also recorded as cleanly as possible. However, it is a time consuming process to capture clean audio and manually author metadata. The use-cases described below demonstrate how our system has the potential to overcome these limitations of traditional audio workflows. In addition to the clean audio recordings, a number of "live" takes were recorded. As the actors were in this case allowed to overlap their lines, the sound designer commented that they could "really tell...how much [the semi-anechoic recording] process affected acting". This suggests that automatic encoding of the actors' positions might lead to more natural performances from the actors, as well as reducing production effort.

B. Scene Capture With Automated Metadata

The proposed system allows spatial audio content to be captured in an object-based form, and rendered in a format-agnostic way. As an example, consider the scene described in Section III-A, comprising two moving talkers. Through audio-visual tracking by the method of Section III-A2, the talkers' positions over a 20 s sequence were tracked and converted to the object metadata as point sources, forming a two-object scene which was then sent to the renderer over UDP. These positions are considered as the ground-truth positions of the two talkers in the scene. Audio recorded using lapel microphones worn by each talker was also sent to the renderer, and the scene was rendered over a "9 + 10 + 3" setup [68], in addition to a dense circular array of 36 virtual loudspeakers (i.e., having 10° spacing). Finally, binaural feeds were acquired by auralization of the virtual loudspeaker layouts (see Fig. 7). Two cases were considered: ideal rendering using anechoic HRTF measurements of a Neumann KU 100 dummy head [69], and auralization of a listening room using binaural room impulse responses (BRIRs) acquired in the BBC listening room [70], both at the sweet spot and in a second position 62 cm forward and 65 cm left of the sweet spot.

To predict the resulting listening experience, a perceptual model combining ITD and ILD features and an auditory modeling front-end was utilized [71], using the implementation in the Auditory Modeling Toolbox [72]. This model is able to localize multiple concurrent speakers. For each time frame (2.5 s) and frequency band, the model gives as output a histogram of the estimated DOAs. Histograms were averaged over 500–1400 Hz, following [71], and processed by picking at most two prominent peaks in each time frame. Candidate DOAs associated with these peaks were used to update the states of particles, from which the talker trajectories were estimated using particle filtering [73]. Quantitative estimates of localization performance were obtained by taking the root-mean-square-error (RMSE), comparing the trajectories to the ground-truth metadata positions over the 20 s sequence.

Fig. 9 shows the localization results for three cases, each utilizing a 9 + 10 + 3 loudspeaker setup. On each plot, the extracted candidate DOAs are shown, together with the two estimated talker trajectories (thick lines) and the original object metadata values (dashed lines). Metadata positions outside the range ±90° are wrapped to be within the range for plotting, as front-back ambiguities cannot be resolved with ITD and ILD features. Fig. 9(a) shows the case where anechoic HRTFs were used for rendering. Here, it can be seen that the estimated tra-

jectories fit the target positions very well, especially over the first 10 s of the sequence. The RMSE (9.5° averaged over both talkers) over the 20 s sequence is comparable to that of the dense circular loudspeaker array (10.2°). These kinds of virtual loudspeaker setup could be used for headphones or with transaural processing, accurately conveying the perceived metadata while limiting the need to store HRTFs for a full range of potential object positions.

Fig. 9(b) and (c) show the effect of activating the listener tracking in a listening room. Fig. 9(b) shows the estimated localization for an off-center listener, while Fig. 9(c) shows the localization when the listener's position is compensated. In both cases, the listening room BRIRs were used for auralization, which limits the DOA model's performance, especially towards the lateral positions. Nevertheless, over the 20 s sequence, activating the listener compensation gave an RMSE of 12.3°, compared to 20.4° for the uncompensated case (and 11.6° for the listening room sweet spot BRIR).

C. Scene Capture Utilizing Acoustic Scene Segmentation

For the final use-case, we consider two live recording scenarios. A common approach is to mix close microphone object signals with a channel-based capture of the entire scene, facilitating greater editing capability. If close microphone signals are not available (for example, if there is limited setup time or close microphones must be avoided for visual reasons), the methods described in Section III-B can be employed to estimate object signals. These object signals can then be broadcast alongside the channel feeds, to allow a content producer or end-user to modify the overall balance of the scene. This kind of hybrid approach is supported in current standards [8], [11]. We present two examples, utilizing the BSS and beamforming components, respectively. Performances were recorded in a large recording studio (RT60 1.1 s), with a 48 channel microphone array in addition to a variety of spatial microphone techniques (documented in [74]). The listening tests reported below were conducted using a standardized "0 +5 + 0" setup [68] with Genelec 8020B loudspeakers in an acoustically-treated listening room with RT60 conforming to the ITU recommendation [75] above 400 Hz.

1) Blind Source Separation: To investigate the utility of BSS to enable object-based remix of stereo speech content, subjective and objective perceptual experiments were conducted. Different TIMIT sentences spoken simultaneously by two talkers were recorded with a pair of high-quality omnidirectional microphones, 18 cm apart, approximately 4 m from the talkers. Lapel microphone signals were also recorded, to provide close reference signals for the objective evaluation. In the stereo recording, one talker was 4.6 dB louder than the other, according to the estimated signal-to-interference ratios (SIRs) calculated by BSS Eval [76]. Mandel's method [35] was used to estimate the quieter talker as an object, with a view to allowing a producer or listener to adjust the level of that talker to make the associated speech clearer. The extracted speech object had 3.54 dB SIR.

In the subjective experiment, listeners were presented with the reference stereo recording (left and right signals rendered directly to ±30 °) and the BSS-estimated object. They were asked to *"adjust the slider [controlling the extracted object level] until the target talker is as clear and easy to understand as possible, whilst ensuring that the overall audio quality remains at an acceptable level (compared to the reference)."* The BSS object was rendered at azimuths $\{0, 15, 30°\}$, with three repeats, giving nine ratings per listener. Additionally, a threshold of audibility was determined: listeners were presented with the same stimulus (object at 0°) and asked to *"adjust the [object] level to the point immediately before the mix is different to the reference."* This part also included three repeats. Ten experienced listeners completed the tests, of whom seven were native English speakers. In a post-screening of the data, the results of one participant were removed as they were found to give inconsistent threshold judgments; the remaining threshold judgements were normally distributed. The results are shown as horizontal error bars in Fig. 10. The mean mixing level averaged over azimuth (0.2 dB relative to the reference) differed significantly from the threshold of audibility (−14.9 dB) according to a two-sample t-test ($t = 9.73$, $p < 0.01$). This shows the benefit of BSS; there is a region in which the BSS-extracted object is audible and makes the target talker clearer while maintaining acceptable quality. An analysis of variance (ANOVA) showed no significant effects of azimuth ($F = 0.85$, $p = 0.43$) or repeat ($F = 0.98$, $p = 0.38$) on the acceptability threshold.

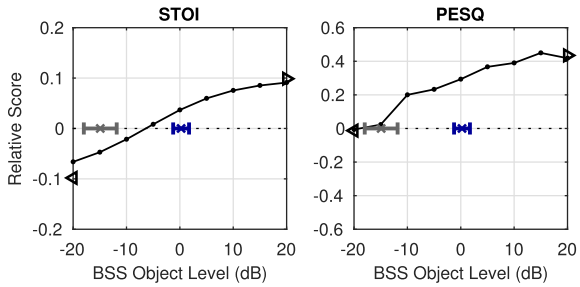


Fig. 10. Relative objective STOI (left) and PESQ (right) scores as a function of level, for an object extracted by BSS mixed into a stereo recording, with the mixture-only (◁) and object-only (▷) scores. Perceptual thresholds of audibility (grey) and acceptability (blue) with 95% confidence intervals are also shown.

The objective evaluation employed metrics of short-time objective intelligibility (STOI [77]), which predicts speech intelligibility, and PESQ, which predicts speech quality. The mono sum of the stereo reference, mixed with the extracted speech object at relative levels in the range ±20 dB, was presented to the models. Prior to processing, all signals were downsampled to 16 kHz and each test mixture was loudness matched to have the same loudness as the reference lapel microphone signal. Objective scores were calculated as the average over those for sentence-level clips in the recording (4 clips with average duration 3.2 s for the target talker; 5 clips with average duration 2.7 s for the interfering talker). After BSS, the extracted object had STOI and PESQ scores of 0.44 and 1.87, respectively. Nevertheless, the judgement about whether one talker is clearer than another depends on the relative mix of both talkers. Therefore relative objective scores were calculated (target talker score − interfering talker score) and are plotted as curves in Fig. 10. The −0.1 relative STOI score for the target talker in the original stereo recording (SIR −0.48 dB) confirms that the interfering talker is more intelligible than the target talker before mixing the extracted object into the scene. By increasing the object's level

in the mixture, the relative STOI and PESQ scores increased. At the mean mixing level determined in the subjective tests, the relative scores are both positive, confirming that introducing the separated speech into the mix has resulted in an enhancement.

2) Beamforming: A jazz group (piano, guitar, bass guitar, drums), arranged roughly along an arc approximately 4 m from the 48-channel microphone array, was also recorded. A 5-channel baseline mix was produced [74, Sec. 4]. In the resulting mix, the piano had a lower acoustic level than the other instruments, and was distant and lacked definition. Consequently, in order to improve the overall mix, the microphone array was steered towards the piano to extract an object signal. The beamformer applied was a frequency-independent 9th order hypercardioid, created by least-squares matching of the beampatterns while constraining the white noise gain not to move below 10 dB. The extracted object signal was bandlimited using two parametric equalization sections with gains of $-23.7/-17.0$ dB, center frequencies 231/6695 Hz, and Q factors 3.8/0.5, giving upper and lower -6 dB points at 325 Hz and 5 kHz, respectively, approximately corresponding to the effective frequency range of the beamformer.

A perceptual test was conducted to evaluate the effect of the extracted object in the context of the overall scene. The channels from the reference 5-channel mix were played as point source objects located at the loudspeaker positions, and the piano object was mixed into the scene with six different levels (-4, -2, 0, 2, 4, and 6 dB relative to the reference) and two positions (10 degrees—corresponding to the approximate position of the piano in the reference scene—and 25 degrees). Stimuli were loudness-matched using the meter described in Section IV-B [49]. The thirteen stimuli (six levels $\times$ two positions, plus a hidden reference, which was identical to the explicit reference) were presented to the listeners on a multiple stimulus interface, and listeners were asked to rate their preference for the test stimuli compared to the explicit reference (where the mid-point of the scale was no preference). Twelve experienced listeners—of whom four had extensive experience creating 3D audio mixes—completed the tests. To compensate for different use of the scale by participants, the scores were normalized by dividing all scores by the standard deviation of the scores [78]. The results from two participants were discarded (one participant was shown to be inconsistent; the other misidentified the hidden reference).

An ANOVA was performed, showing a significant effect of beamformer object level on preference ($F = 29.1, p < 0.01$). The results are shown in Fig. 11, which shows that an increase in preference was given when the beamformer object level was -2 dB (at 25 degrees) or -4 dB (at both angles). In these cases, the object helped to make the piano less distant in the mix without affecting the overall quality. On the other hand, as the mix ratio increased, artifacts due to the beamforming (namely band limiting and the effect of the interfering sources altering the spatial image) were more exposed. There was a small, non-significant increase in preference for the piano object rendered at 25 degrees compared to 10 degrees. Comments recorded from participants suggested that this had the effect of widening the perceived piano image. In summary, subtle addition of the separated object was shown to be preferred by the listeners to just the channel-based mix, even where no close microphone signals were available.

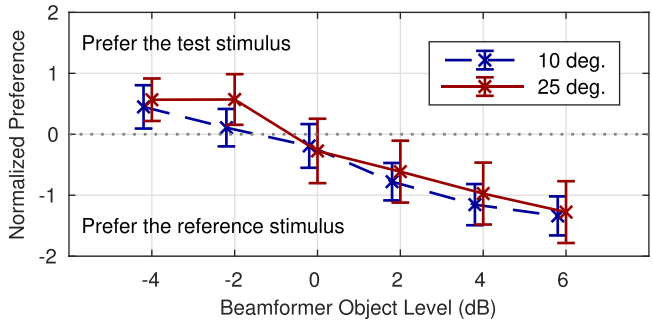


Fig. 11. Listener preference as a function of level and object position, for a jazz group recording including a piano object extracted by beamforming.

D. System Capabilities, Limitations and Future Opportunities

State-of-the-art object-based audio workflows require clean audio, manually authored metadata, and listeners positioned in the sweet spot. Our proposed system overcomes these limitations by the inclusion of audio-visual interfaces for metadata capture and listener-tracked rendering, and novel applications of BSS and beamforming signal processing for use in clean audio acquisition. Furthermore, the integration of state-of-the-art components means that our system is able, for example, to deliver an accurate perception of moving talkers for a listener outside of the sweet spot.

We demonstrated in Sections VII-B and VII-C that the proposed objectification components can be used to capture object audio and metadata, even where no close microphone signals are available. The producer thus has greater control over the scene than for recordings made with traditional channel-based techniques. Our subjective evaluations showed that this approach to capture and editing improved the resulting listener experience. An ongoing challenge for object-based audio segmentation is to estimate clean, high quality, objects. Informally, we observed amplitude panning effects when attempting to re-spatialize extracted objects with imperfect separation. This effectively imposes an upper limit on the spatial remixing achievable for objects captured in this way. In addition, there is a need for production tools which incoporate objectification.

The object-based representation underpins the whole end-to-end system. Of the object types currently in our model (see Section V), point source and plane wave objects are the most commonly used. Our objectificaton techniques and production tools do not yet support the other, more experimental, object types. Nevertheless, these object types, combined with new descriptive or semantic fields, may support the producer to encode their artistic intent into the scene. Future object-based workflows should support multiple object types across the whole end-to-end pipeline.

The renderer is of vital importance for object-based scenes. Although the VBAP rendering is said to be format-agnostic, in practice performance depends on the available loudspeakers. For instance, the estimated RMSE in localization performance for an ideal 5 channel system rendering the 20 s scene from

Section VII-B was 19.6°, averaged over both objects. The main cause of this degradation is that the scene contains objects in positions where the loudspeakers are not sufficiently close together for amplitude panning to be effective (i.e., behind the listener). Future object renderers may be able to account for such limitations, as well as to be able to convincingly render objects with extent, diffuseness, and varying distance, that have been captured by the other future system components.

Finally, the transition from channel-based to object-based representations of audio raises a number of interesting questions about how listeners experience the content. In particular, future work should investigate how listeners would use controls to interact with the sound scene, and what kinds of controls they would like to have.

VIII. Summary

In this paper, we proposed an audio-visual system for spatial audio, covering the full pipeline of audio production from capture of acoustic signals to monitoring by a perceptual meter or listener. The proposed system has a novel architecture, including audio-visual interfaces for capture and rendering and a novel component for direct capture of content into an object-based representation. We propose a new object metadata scheme and describe the design and implementation of an open, flexible rendering architecture. A discussion of the system's capabilities was formed around three end-to-end use-cases: production of radio drama scenes; scene capture with metadata estimated from moving talkers, rendered over different loudspeaker setups, and evaluated using an objective binaural localization model; and scene capture with audio extracted using BSS (to remix between two talkers) and beamforming (to remix a recording of a jazz group), evaluated in formal listening tests. The latter experiments showed that extracted audio objects can be added to channel-based recordings, thus allowing remixing and respatialization, while maintaining acceptable audio quality. Finally, we discussed future opportunities.

Acknowledgment

The authors are grateful to the wider S3A project team for their input on the work described, in particular Tim Brookes, Bill Davies, Bruno Fazenda, Russell Mason, Ben Shirley and Wenwu Wang. Data underlying the findings and certain system components are available; details are available from `https://doi.org/10.15126/surreydata.00845514`.

References

[1] G. Thomas *et al.*, "State of the art and challenges in media production, broadcast and delivery," in *Media Production, Delivery and Interaction for Platform Independent Systems*. Hoboken, NJ, USA: Wiley, 2013, pp. 5–73.

[2] S. Spors *et al.*, "Spatial sound with loudspeakers and its perception: A review of the current state," *Proc. IEEE*, vol. 101, no. 9, pp. 1920–1938, Sep. 2013.

[3] B. Shirley and R. Oldfield, "Clean audio for TV broadcast: An object-based approach for hearing-impaired viewers," *J. Audio Eng. Soc.*, vol. 63, no. 4, pp. 245–256, 2015.

[4] M. Mann, A. W. Churnside, A. Bonney, and F. Melchior, "Object-based audio applied to football broadcasts," in *Proc. ACM Int. Workshop Immersive Media Exp.*, Barcelona, Spain, 2013, pp. 13–16.

[5] M. Evans *et al.*, "Creating object-based experiences in the real world," in *Proc. IBC Conf.*, Amsterdam, The Netherlands, Sep. 2016, p. 34.

[6] M. Armstrong *et al.*, "Object-based broadcasting—Curation, responsiveness and user experience," in *Proc. IBC Conf.*, Amsterdam, The Netherlands, Sep. 2014, p. 12.

[7] B. Shirley, R. Oldfield, F. Melchior, and J.-M. Batke, "Platform independent audio," in *Media Production, Delivery and Interaction for Platform Independent Systems*. Hoboken, NJ, USA: Wiley, 2013, pp. 130–165.

[8] 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, Aug. 2015.

[9] ITU-R, "Recommendation BS.2076-0: Audio Definition Model," Int. Telecommun. Union, Geneva, Switzerland, 2015.

[10] *MDA: Object-Based Audio Immersive Sound Metadata and Bitstream*, Std. ETSI-TS-103-223, Apr. 2015.

[11] *Digital Audio Compression (AC-4) Standard Part 2: Immersive and Personalized Audio*, Std. ETSI-TS-103-190-2, Sep. 2015.

[12] *AC-4 Object Audio Renderer for Consumer Use*, Std. ETSI-TS-103-448, Sep. 2016.

[13] M. Weitnauer *et al.*, "D2.2: Interim reference architecture specification and integration report," ORPHEUS public deliverable, Tech. Rep., 2017. [Online]. Available: https://orpheus-audio.eu/public-deliverables/.

[14] F. Rumsey, *Spatial Audio*. Waltham, MA, USA: Focal Press, 2001.

[15] F. Rumsey and T. McCormick, *Sound and Recording: An Introduction*. New York, NY, USA: Taylor & Francis, 2006.

[16] G. T. Heineman and W. T. Councill, *Component-Based Software Engineering: Putting the Pieces Together*. Boston, MA, USA: Addison-Wesley, 2001.

[17] IETF, "RFC 7159—The JavaScript Object Notation (JSON) data interchange format," Internet Engineering Task Force, RFC, 2014. [Online]. Available: https://tools.ietf.org/html/rfc7159

[18] E. Corteel *et al.*, "An open 3D audio production chain proposed by the Edison 3D project," in *Proc. 140 Conv. Audio Eng. Soc.*, Paris, France, Jun. 2016, Paper 9589.

[19] C. Pike, R. Taylor, T. Parnell, and F. Melchior, "Object-based 3D audio production for virtual reality using the audio definition model," in *Proc. AES Int. Conf. Audio Virtual Augmented Reality*, Los Angeles, CA, USA, Sep. 2016, Paper 2-1.

[20] R. Oldfield, B. Shirley, and N. Cullen, "Demo paper: Audio object extraction for live sports broadcast," in *Proc. Int. Conf. Multimedia Expo Workshops*, San Jose, CA, USA, Jul. 2013, pp. 1–2.

[21] R. Oldfield, B. Shirley, and J. Spille, "Object-based audio for interactive football broadcast," *Multimedia Tools Appl.*, vol. 74, no. 8, pp. 2717–2741, 2015.

[22] D. Comminiello *et al.*, "Intelligent acoustic interfaces with multisensor acquisition for immersive reproduction," *IEEE Trans. Multimedia*, vol. 17, no. 8, pp. 1262–1272, Aug. 2015.

[23] M. Ye *et al.*, "A survey on human motion analysis from depth data," in *Time-of-Flight and Depth Imaging Sensors. Algorithms, and Applications*, vol. 8200. Berlin, Germany: Springer, 2013, pp. 149–187.

[24] C. Redondo-Cabrera, R. Lopez-Sastre, and T. Tuytelaars, "All together now: Simultaneous detection and continuous pose estimation using a Hough forest with probabilistic locally enhanced voting," in *Proc. Brit. Mach. Vision Conf.*, Nottingham, U.K., Sep. 2014, pp. 1–12.

[25] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. Van Gool, "Random forests for real time 3D face analysis," *Int. J. Comput. Vision*, vol. 101, no. 3, pp. 437–458, Feb. 2013.

[26] J. M. Saragih, S. Lucey, and J. F. Cohn, "Face alignment through subspace constrained mean-shifts," in *Proc. 12th Int. Conf. Comput. Vision*, Kyoto, Japan, Sep. 2009, pp. 1034–1041.

[27] M. Barnard *et al.*, "Robust multi-speaker tracking via dictionary learning and identity modeling," *IEEE Trans. Multimedia*, vol. 16, no. 3, pp. 864–880, Apr. 2014.

[28] A. K. Katsaggelos, S. Bahaadini, and R. Molina, "Audiovisual fusion: Challenges and new approaches," *Proc. IEEE*, vol. 103, no. 9, pp. 1635–1653, Sep. 2015.

[29] R. P. S. Mahler, "Multitarget Bayes filtering via first-order multitarget moments," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 39, no. 4, pp. 1152–1178, Oct. 2003.

[30] Q. Liu, T. de Campos, W. Wang, and A. Hilton, "Identity association using PHD filters in multiple head tracking with depth sensors," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, Shanghai, China, Mar. 2016, pp. 1506–1510.

[31] C. Knapp and G. C. Carter, "The generalized correlation method for estimation of time delay," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. ASSP-24, no. 4, pp. 320–327, Aug. 1976.

[32] A. G. Westner, "Object-based audio capture: Separating acoustically-mixed sounds," Master's thesis, Massachusetts Inst. Technol., Cambridge, MA, USA, Feb. 1999.
[33] P. Comon, "Independent component analysis, A new concept?" *Signal Process.*, vol. 36, no. 3, pp. 287–314, Apr. 1994.
[34] D. Wang, "On ideal binary mask as the computational goal of auditory scene analysis," in *Speech Separation by Humans and Machines*, P. Divenyi, Ed. New York, NY, USA: Springer, 2005, ch. 12, pp. 181–197.
[35] M. I. Mandel, R. J. Weiss, and D. Ellis, "Model-based expectation-maximization source separation and localization," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 2, pp. 382–394, Feb. 2010.
[36] A. Ozerov and C. Févotte, "Multichannel nonnegative matrix factorization in convolutive mixtures for audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 18, no. 3, pp. 550–563, Mar. 2010.
[37] A. A. Nugraha, A. Liutkus, and E. Vincent, "Multichannel audio source separation with deep neural networks," *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, vol. 24, no. 9, pp. 1652–1664, Sep. 2016.
[38] A. W. Rix, J. G. Beerends, M. P. Hollier, and A. P. Hekstra, "Perceptual evaluation of speech quality (PESQ)—A new method for speech quality assessment of telephone networks and codecs," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, vol. 2, Salt Lake City, UT, USA, May 2001, pp. 749–752.
[39] Q. Liu, W. Wang, P. J. B. Jackson, and T. J. Cox, "A source separation evaluation method in object-based spatial audio," in *Proc. 23rd Eur. Signal Process. Conf.*, Nice, France, Aug. 2015, pp. 1088–1092.
[40] B. D. VanVeen and K. M. Buckley, "Beamforming: A versatile approach to spatial filtering," *IEEE Acoust., Speech Signal Process. Mag.*, vol. 5, no. 2, pp. 4–24, Apr. 1988.
[41] Y. Huang, J. Chen, and J. Benesty, "Immersive audio schemes," *IEEE Signal Process. Mag.*, vol. 28, no. 1, pp. 20–32, Jan. 2011.
[42] K. Kumatani, J. McDonough, and B. Raj, "Microphone array processing for distant speech recognition: From close-talking microphones to far-field sensors," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 127–140, Nov. 2012.
[43] P. Coleman, P. J. B. Jackson, and J. Francombe, "Audio object separation using microphone array beamforming," in *Proc. 138 Conv. Audio Eng. Soc.*, Warsaw, Poland, May 2015, Paper 9296.
[44] K. Kinoshita *et al.*, "A summary of the reverb challenge: State-of-the-art and remaining challenges in reverberant speech processing research," *EURASIP J. Adv. Signal Process.*, vol. 2016, no. 1, pp. 1–19, 2016.
[45] *AES Recommended Practice for Digital Audio Engineering—Serial Multichannel Audio Digital Interface (MADI)*, Std. AES10-1991, 1991.
[46] Y. Tang, M. Cooke, B. M. Fazenda, and T. J. Cox, "A glimpse-based approach for predicting binaural intelligibility with single and multiple maskers in anechoic conditions," in *Proc. Interspeech*, Dresden, Germany, Sep. 2015, pp. 2568–2572.
[47] T. Komori *et al.*, "Subjective loudness of 22.2 multichannel programs," in *Proc. 138 Conv. Audio Eng. Soc.*, Warsaw, Poland, May 2015, Paper 9219.
[48] J. Francombe, T. Brookes, R. Mason, and F. Melchior, "Loudness matching multichannel audio programme material with listeners and predictive models," in *Proc. 139 Conv. Audio Eng. Soc.*, New York, NY, USA, 2015, Paper 9464.
[49] ITU-R, "Recommendation BS.1770-4: Algorithms to measure audio programme loudness and true-peak audio level," Int. Telecommun. Union, Geneva, Switzerland, 2015.
[50] C. Pike and F. Melchior, "An assessment of virtual surround sound systems for headphone listening of 5.1 multichannel audio," in *Proc. 134 Conv. Audio Eng. Soc.*, Rome, Italy, May 2013, Paper 8819.
[51] M. Geier, J. Ahrens, and S. Spors, "Object-based audio reproduction and the audio scene description format," *Org. Sound*, vol. 15, no. 3, pp. 219–227, Dec. 2010.
[52] N. Peters, T. Lossius, and J. C. Schacher, "The spatial sound description interchange format: Principles, specification, and examples," *Comput. Music J.*, vol. 37, no. 1, pp. 11–22, 2013.
[53] P. Coleman *et al.*, "Object-based reverberation for spatial audio," *J. Audio Eng. Soc.*, vol. 65, no. 1/2, pp. 66–77, 2017.
[54] S. Füg *et al.*, "Design, coding and processing of metadata for object-based interactive audio," in *Proc. 137 Conv. Audio Eng. Soc.*, Los Angeles, CA, USA, 2014, Paper 9097.
[55] M. Geier, T. Hohn, and S. Spors, "An open-source C++ framework for multithreaded realtime multichannel audio applications," in *Proc. Linux Audio Conf.*, Stanford, CA, USA, Apr. 2012, pp. 183–188.
[56] F. Zotter and M. Frank, "All-round ambisonic panning and decoding," *J. Audio Eng. Soc.*, vol. 60, no. 10, pp. 807–820, 2012.
[57] V. Pulkki, "Virtual sound source positioning using vector base amplitude panning," *J. Audio Eng. Soc.*, vol. 45, no. 6, pp. 456–466, 1997.
[58] M. F. Simón Gálvez, D. Menzies, F. M. Fazi, T. de Campos, and A. Hilton, "A listener position adaptive stereo system for object-based reproduction," in *Proc. 138 Conv. Audio Eng. Soc.*, Warsaw, Poland, 2015, Paper 9246.
[59] M. F. Simón Gálvez, D. Menzies, R. Mason, and F. M. Fazi, "Object-based audio reproduction using a listener-position adaptive stereo system," *J. Audio Eng. Soc.*, vol. 64, no. 10, pp. 740–751, 2016.
[60] C. Pike, F. Melchior, and A. Tew, "Descriptive analysis of binaural rendering with virtual loudspeakers using a rate-all-that-apply approach," in *Proc. Int. Conf. Headphone Technol.*, Aalborg, Denmark, Aug. 2016, Paper 5–3.
[61] A. Mouchtaris, P. Reveliotis, and C. Kyriakakis, "Inverse filter design for immersive audio rendering over loudspeakers," *IEEE Trans. Multimedia*, vol. 2, no. 2, pp. 77–87, Jun. 2000.
[62] O. Kirkeby, P. A. Nelson, and H. Hamada, "Local sound field reproduction using two closely spaced loudspeakers," *J. Acoust. Soc. Amer.*, vol. 104, no. 4, pp. 1973–1981, 1998.
[63] T. Takeuchi and P. A. Nelson, "Optimal source distribution for binaural synthesis over loudspeakers," *J. Acoust. Soc. Amer.*, vol. 112, no. 6, pp. 2786–2797, 2002.
[64] M.-S. Song, C. Zhang, D. Florencio, and H.-G. Kang, "An interactive 3-D audio system with loudspeakers," *IEEE Trans. Multimedia*, vol. 13, no. 5, pp. 844–855, Oct. 2011.
[65] M. F. Simón Gálvez and F. M. Fazi, "Sweet-spot-independent binaural reproduction with a listener-adaptive loudspeaker array," in *Proc. 22nd Int. Congr. Acoust.*, Buenos Aires, Argentina, 2016, Paper ICA2016-598, pp 1–10.
[66] M. F. Simón Gálvez, T. Takeuchi, and F. M. Fazi, "A listener adaptive optimal source distribution system for virtual sound imaging," in *Proc. 140 Conv. Audio Eng. Soc.*, Paris, France, Jun. 2016, Paper 9574.
[67] J. Woodcock *et al.*, "Presenting the S3A object-based audio drama dataset," in *Proc. 140 Conv. Audio Eng. Soc.*, Paris, France, Jun. 2016, Paper 255.
[68] ITU-R, "Recommendation ITU-R BS.2051-0: Advanced sound system for programme reproduction," Int. Telecommun. Union, Geneva, Switzerland, 2014.
[69] B. Bernschütz, "A spherical far field HRIR/HRTF compilation of the Neumann KU 100," in *Proc. 40th Italian Annu. Conf. Acoust. /39th German Annu. Conf. Acoust.*, 2013, p. 29.
[70] T. Nixon, A. Bonny, and F. Melchior, "A reference listening room for 3D audio research," in *Proc. 3rd Int. Conf. Spatial Audio*, Graz, Austria, Sep. 2015, Paper 016, pp 1–6.
[71] M. Dietz, S. D. Ewert, and V. Hohmann, "Auditory model based direction estimation of concurrent speakers from binaural signals," *Speech Commun.*, vol. 53, no. 5, pp. 592–605, 2011.
[72] P. Søndergaard, and P. Majdak, "The auditory modeling toolbox," in *The Technology of Binaural Listening*, J. Blauert, Ed. Berlin, Germany: Springer, 2013, pp. 33–56.
[73] A. Doucet, S. Godsill, and C. Andrieu, "On sequential Monte Carlo sampling methods for Bayesian filtering," *Stat. Comput.*, vol. 10, no. 3, pp. 197–208, 2000.
[74] J. Francombe *et al.*, "Production and reproduction of program material for a variety of spatial audio formats," in *Proc. 138 Conv. Audio Eng. Soc.*, Warsaw, Poland, 2015, Paper 199.
[75] ITU-R, "Recommendation ITU-R BS.1116-3, Methods for the subjective assessment of small impairments in audio systems," Int. Telecommun. Union, Geneva, Switzerland, Feb. 2015.
[76] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 4, pp. 1462–1469, Jul. 2006.
[77] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An algorithm for intelligibility prediction of time–frequency weighted noisy speech," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 7, pp. 2125–2136, Sep. 2011.
[78] 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, no. 3, pp. 212–225, Mar. 2017.

Authors' photographs and biographies not available at the time of publication.