Framework for Web Delivery of Immersive Audio Experiences Using Device Orchestration

Kristian Hentschel
BBC R&D
MediaCityUK, Salford, UK
kristian.hentschel@bbc.co.uk

Jon Francombe
BBC R&D
MediaCityUK, Salford, UK
jon.francombe@bbc.co.uk

The Adjunct Proceedings of ACM TVX 2019, Manchester UK, June 2019.

Abstract
This demonstration introduces the use of orchestrated media devices and object-based broadcasting to create immersive spatial audio experiences. Mobile phones, tablets, and laptops are synchronised to a common media timeline and contribute one or more individually delivered audio objects to the overall mix. A rule set for assigning objects to devices was developed through a trial production—a 13-minute audio drama called *The Vostok-K Incident*. The same timing model as in HbbTV2.0 media synchronisation is used, and future work could augment linear television broadcasts or create novel interactive audio-visual experiences for multiple users. The demonstration will allow delegates to connect their mobile phones to the system. A unique mix is created based on the number and selected locations of connected devices.

Author Keywords
Device orchestration; immersive audio; object-based broadcasting.

ACM Classification Keywords
H.5.5 [Sound and Music Computing]: Systems; H.5.3 [Group and Organization Interfaces]: Collaborative Computing; C.2.4 [Distributed Systems]: Distributed Applications

Audio in object-based broadcasting

Object-based broadcasting is the idea of storing the media assets that constitute a programme alongside metadata that describe how those assets should be assembled. This enables flexible rendering; for example, content might be modified to suit the reproduction device, the user preferences, or for interactive or variable length applications. In object-based audio, the objects are the different sound sources from the mix, and the metadata generally include details such as position and level (but may also include more detailed semantic descriptions [10]).

Introduction

Media experiences are increasingly taking place on mobile devices—either as the main device or as a secondary device being used at the same time. The Internet of Things offers ever more connectable devices, many with media playback capabilities. Device orchestration is the concept of using these diverse devices to create or augment an immersive media experience.

Current immersive audio experiences in the home rely on channel-based delivery matched to common reproduction systems, such as stereo or 5.1 surround loudspeaker configurations. However, only 11.5% of users in the UK report having a surround system, according to a recent survey by Cieciura *et al.* [1]. Alternatively, headphones can be used to reproduce binaural audio, but this is limited to a single listener, and wearing headphones is not always appropriate.

This demonstration introduces a new approach that uses a web-based framework alongside object-based audio (see sidebar) to deliver immersive audio with commodity mobile devices. Listeners can connect their mobile phones and select a location; their device becomes part of the sound system and plays appropriate content. A trial production (a short audio drama) was used to motivate the development of the system, providing requirements for an orchestration ruleset, accuracy of synchronisation, and the user interface.

Prior work

Synchronised second-screen applications for television broadcasts have been enabled by the development of Hybrid Broadcast Broadband TV (HbbTV 2.0). Previous demonstrations focused on textual or visual content, although alternative audio tracks (such as audio description or director's commentary) have also been shown [9].

Mobile phones have been used for orchestrated audio reproduction before, with demonstrations largely falling into two categories: (i) in interactive performance scenarios where individual sounds are triggered by each user [8]; and (ii) for reproduction of channel-based spatial audio mixes requiring very accurate synchronisation between loudspeakers [5]. However, object-based broadcasting enables a new way to create immersive experiences by using synchronised devices to reproduce different aspects of a scene. This approach has performed well in perceptual tests; the listening experience achieved with small loudspeakers in non-standard locations was shown to be comparable to a surround sound system by Woodcock *et al.* [11].

Web framework for audio device orchestration

An orchestrated immersive audio scene consists of individual audio assets synchronously played back from multiple devices. The prototype system architecture for achieving this is outlined in Figure 1. A single main device creates a session: additional devices can join by entering a pairing code and selecting a location zone. All devices maintain a wall clock synchronised to a central server. Metadata describe timings and placement constraints for every object. The main device runs the placement algorithm and assigns objects to devices. Each device requests only the audio objects assigned to it, and dynamically mixes them on the device.

The web platform was selected for prototyping the experience because it is easily accessible to a wide audience. WebSockets are used for asynchronous communication including wall clock synchronisation. The Web Audio API is used for rendering multiple synchronised audio streams. A user interface template for pairing devices and controlling

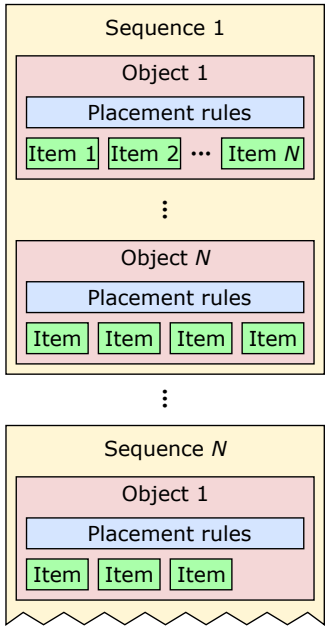

Figure 2: A *sequence* metadata file describes how *objects* should be distributed to available devices, and when each *rendering item* should be played.

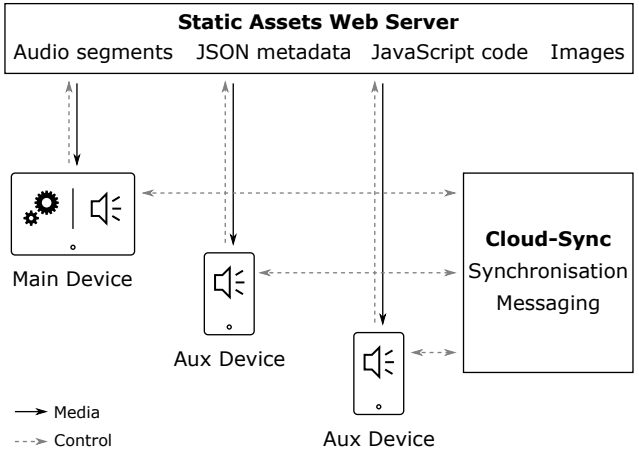

Figure 1: System diagram; one *main* and two *auxiliary* devices request audio assets and metadata, synchronise to a server, and exchange messages. The main device decides which audio objects should be played by each device.

playback was written in JavaScript using the React framework and can be customised for specific productions.

Devices exchange messages with a server to measure the round trip time and determine their hardware clock's offset and speed difference. Multiple clock synchronisation libraries are available [6, 4]. The Cloud-Sync solution developed in the 2-IMMERSE project [7] was chosen because it directly corresponds to the HbbTV2.0 timeline model and offers a scalable server implementation. A synchronised wall clock is used to derive media timelines for every individually scheduled audio object. The service also provides messaging between devices in a session and notifications when a device joins or leaves.

The content metadata hierarchy is illustrated in Figure 2. An object-based audio mix is exported from a digital audio workstation (DAW) with a full-length mono audio file for each *object*. Most objects are silent most of the time, so the silence is removed and the objects are split into *rendering items*. An experience may transition between multiple continuous *sequences* i.e., different parts or sections of the content. Each is defined by a JavaScript Object Notation (JSON) metadata file describing the objects and placement rules, and the start times and download locations for individual rendering items. Long running rendering items are delivered using Dynamic Adaptive Streaming over HTTP (DASH). The browser on each device downloads, plays, and combines the rendering items for objects assigned to it using the Web Audio API. Additional effects, such as dynamics compression, may be applied here. The metadata hierarchy is illustrated in Figure 2.

The main device maintains a list of auxiliary devices in the current session. It runs a placement algorithm to assign a combination of object identifiers to each device when a device joins, leaves, or changes location. The assignments are distributed to all devices in the session through a publish-subscribe message broker included in the Cloud-Sync service. Each device only downloads the *rendering items* for those *objects* currently assigned to it, shortly before they are played out.

A user interface template for starting sessions, connecting devices, selecting their location, controlling the audio playback, and transitioning between different content sequences was also created. This shows how the framework can be integrated with React to build a responsive web application. The template was customised and extended to create the interface for a trial production, discussed in the following section.

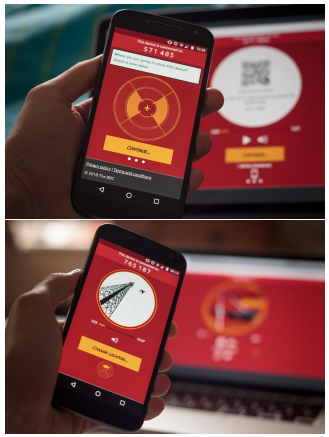

Figure 3: *The Vostok-K Incident* was released on BBC Taster in September 2018. The photographs above show the user interface: after pairing via a QR code or link and numerical code, an approximate location is selected (top); then each device displays artwork related to the audio objects it has been assigned (bottom).

Demonstration

A trial production, the 13-minute audio drama *The Vostok-K Incident*, motivated the development of the framework. Its main story sequence consists of a stereo bed and 61 additional objects: character dialogue, musical elements, ambient sound, and sound effects. The experience begins with a 30-second *loading loop*, so users can connect their devices without missing the beginning of the story.

Attendees may connect their own devices to augment the stereo reproduction, and listen to a section of the drama. The main screen shows connection instructions and transport controls. Devices are paired using a six-digit code, and a location is selected from a stylised map. The audio mix adapts to the number and location of devices. Artwork related to the objects is shown on each device.

The object-based trial production [3] used regular radio drama recording and sound design techniques, and was initially mixed in stereo. A prototype production system connected to the digital audio workstation (DAW) allowed the producers to audition the effect of different loudspeaker layouts and metadata rules during the immersive mix. For this drama, six different rules were used to allocate objects to devices. The rules covered:

- the location from which objects should be played (for example, some objects might only be allowed into loudspeakers in front of the listener);
- the type of loudspeakers from which objects should be played (for example, some objects were only reproduced if devices were connected whilst others would fall back into the stereo bed); and
- the relationships between objects (for example, some objects would only play if others were not playing, and some objects precluded any others from being played from the same device).

Evaluation and future work

The flexible framework for delivering immersive audio to orchestrated devices created here could be extended in multiple ways. It could integrate visual content, played back on the Web or a smart TV; it could make use of Internet-connected lights and similar devices for additional effects; and it could allow for interactive exploration of non-linear content.

The Vostok-K Incident was released on *BBC Taster*, BBC R&D's public-facing trial platform, and was accessed 3121 times. A questionnaire completed by 210 users indicated that the experience was positive (four out of five stars) and that the framework was successful (only 15% rated 'the set up' as the worst thing about the experience). Interaction logs for 2174 sessions were analysed further by Francombe *et al.* [2]. Preliminary results show that about 65% of sessions that reached the first minute of the main sequence had at least one device connected, and that almost 20% of those that listened to the whole piece connected three or more devices.

Acknowledgements

The authors would like to acknowledge the work of members of *The Vostok-K Incident* production team: James Woodcock and Richard Hughes (University of Salford), and Eloise Whitmore, Tony Churnside, and Ed Sellek (Naked Productions). The production was supported by the EPSRC Programme Grant S3A: Future Spatial Audio for an Immersive Listener Experience at Home (EP/L000539/1).

They would also like to thank Rajiv Ramdhany (BBC R&D) for his assistance in integrating the Cloud-Sync service developed as part of the 2-IMMERSE project.

REFERENCES

1. Craig Cieciura, Russell Mason, Philip Coleman, and Matthew Paradis. 2018. Survey of Media Device Ownership, Media Service Usage, and Group Media Consumption in UK Households. In *Audio Engineering Society Convention 145 (eBrief 456)*. New York, NY, USA.
2. Jon Francombe and Kristian Hentschel. 2019. Evaluation of an immersive audio experience using questionnaire and interaction data. In *23rd International Congress on Acoustics*. Aachen, Germany. Accepted for publication in September 2019.
3. Jon Francombe, James Woodcock, Richard Hughes, Kristian Hentschel, Eloise Whitmore, and Tony Churnside. 2018. Producing audio drama content for an array of orchestrated personal devices. In *Audio Engineering Society Convention 145 (eBrief 461)*. New York, NY, USA.
4. Matt Hammond. 2018. Open-source DVB CSS libraries available. (2018). `https://2immerse.eu/open-source-dvb-css-libraries-available/`
5. Hyosu Kim, SangJeong Lee, Jung-Woo Choi, Hwidong Bae, Jiyeon Lee, Junehwa Song, and Insik Shin. 2014. Mobile maestro: enabling immersive multi-speaker audio applications on commodity mobile devices. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp '14 Adjunct*. `DOI: http://dx.doi.org/10.1145/2632048.2636077`
6. Jean-Philippe Lambert, Sébastien Robaszkiewicz, and Norbert Schnell. 2016. Synchronisation for Distributed Audio Rendering over Heterogeneous Devices, in HTML5. In *Proceedings of the 2nd Web Audio Conference (WAC-2016)*. Atlanta, GA, United States.
7. Rajiv Ramdhany and Christoph Ziegler. 2019. Cloud-Sync Media Synchronisation Service. (2019). `https://github.com/2-IMMERSE/cloud-sync`
8. Sébastien Robaszkiewicz and Norbert Schnell. 2015. Soundworks–A playground for artists and developers to create collaborative mobile web performances. In *Proceedings of the 1st Web Audio Conference (WAC-2015)*. Paris, France.
9. Vinoba Vinayagamoorthy, Rajiv Ramdhany, and Matt Hammond. 2016. Enabling Frame-Accurate Synchronised Companion Screen Experiences. In *Proceedings of the ACM International Conference on Interactive Experiences for TV and Online Video*. New York, NY, USA. `DOI: http://dx.doi.org/10.1145/2932206.2932214`
10. James Woodcock, Jon Francombe, Andreas Franck, Philip Coleman, Richard Hughes, Hansung Kim, Qingju Liu, Dylan Menzies, Marcos F Simón Gálvez, Yan Tang, Tim Brookes, William J. Davies, Bruno M. Fazenda, Russell Mason, Trevor J. Cox, Filippo Maria Fazi, Phiip J. B. Jackson, Chris Pike, and Adrian Hilton. 2018a. A Framework for Intelligent Metadata Adaptation in Object-Based Audio. In *AES International Conference on Spatial Reproduction - Aesthetics and Science*. Tokyo, Japan.
11. James Woodcock, Jon Francombe, Richard Hughes, Russell Mason, William J Davies, and Trevor J Cox. 2018b. A quantitative evaluation of media device orchestration for immersive spatial audio reproduction. In *AES International Conference on Spatial Reproduction - Aesthetics and Science*. Tokyo, Japan.