Research & Development White Paper

WHP 359

October 2019

Producing audio drama content for an array of orchestrated personal devices

J Francombe, J Woodcock, RJ Hughes, K Hentschel, E Whitmore *and* A Churnside

Design & Engineering

BRITISH BROADCASTING CORPORATION

BBC Research & Development
White Paper WHP 359

Producing audio drama content for an array of orchestrated personal devices

Jon Francombe James Woodcock Richard J. Hughes Kristian Hentschel
Eloise Whitmore Anthony Churnside

Abstract

Personal devices with loudspeakers can be orchestrated to increase immersion from low channel count reproduction systems. A trial production was conducted to investigate the content creation workflow and delivery mechanism for orchestrated devices. The content (a 13-minute science-fiction drama entitled *The Vostok-K Incident*) included: a stereo bed; elements only replayed from auxiliary devices; and elements that could either be in the stereo bed or replayed from auxiliary devices. A bespoke production environment was established, including plug-ins for authoring the metadata needed to utilize the rendering ruleset. Ambiguity in the reproduction system, coupled with flexible and complex metadata authoring requirements, made the production challenging and time-consuming. Future work will focus on refining the production process and developing delivery tools.

This paper was presented at the 145th Convention of the Audio Engineering Society, as eBrief number 461. The full published version can be found at http://www.aes.org/e-lib/browse.cfm?elib=19726.

The original paper was published in October 2018; at the time of submission, *The Vostok-K Incident* production had not been released. As of publication of this white paper, *The Vostok-K Incident* is available to view online at https://vostok.virt.ch.bbc.co.uk/.

Additional key words: device orchestration, immersive audio, object-based media

White Papers are distributed freely on request.

Authorisation of the Chief Scientist or Head of Standards is required for publication.

BBC Research & Development
White Paper WHP 359

Producing audio drama content for an array of orchestrated personal devices

Jon Francombe James Woodcock Richard J. Hughes Kristian Hentschel
Eloise Whitmore Anthony Churnside

1 Introduction

It has been shown that *ad hoc* arrays of devices with small, relatively low quality loudspeakers can be orchestrated to augment low channel count reproduction systems and increase the quality of listening experience (particularly due to the increased immersion that can be gained from having loudspeakers at a range of azimuths, elevations, and distances) [1, 2]. This method of spatial audio reproduction is a special case of *media orchestration*, in which the media rendering capabilities of multiple devices can be coordinated to extend the functionality of any single device [3].

Audio experiences using distributed arrays of loudspeakers, including on personal devices, have long been used in compositions and collaborative performances [4]. However, it is only recently that orchestrated devices have been considered for spatial audio reproduction in the home. Research in this area has focused on showing the proof of concept and evaluating the listening experience. Francombe et al. [1] implemented a media device orchestration (MDO) system, augmenting a stereo reproduction with four small consumer loudspeakers. They collected free text responses from a group of listeners and performed a thematic analysis, suggesting that the orchestrated device system performed well on certain perceptual attributes (particularly immersion/envelopment), but less well on technical and content-dependent aspects. For a similar setup in a controlled laboratory evaluation, Woodcock et al. [2] showed that MDO can produce a comparable quality of listening experience to a standard five-channel layout in the sweet spot listening position, and a significantly better experience outside the sweet spot. Whilst validating the concept of using orchestrated personal devices for spatial audio broadcast, the existing studies have not directly addressed some of the challenging technical questions (such as device discovery, pairing, synchronization, localization, and calibration) or content authoring. The content authoring presents a particularly challenging problem, as there is little to no knowledge of the reproduction system at the point of production (in fact, the reproduction is likely to be different in every case).

In order to address some of these issues, a trial audio drama was commissioned. The production was intended to: (i) develop the workflow for creating content for orchestrated devices; and (ii) test the delivery mechanisms and user experience for this type of content. The content creation workflow is the main focus of this paper: the writing and recording (Section 2) and mixing (Section 3) processes are introduced, and challenges are discussed in Section 4. An outlook for future work is presented in Section 5.

2 Concept, script writing, and recording

An original script was commissioned. The writer was aware of the target delivery method (a main stereo bed with an unknown number of auxiliary devices), and was asked to develop the concept considering how the use of auxiliary devices could enhance the storytelling.

The resulting production was a 13-minute science-fiction drama entitled *The Vostok-K Incident*. The drama is set during the Cold War, and takes place in the cockpit of a fighter aircraft. The pilot receives a radio message and is redirected to investigate a mysterious spacecraft. Whilst the

action unfolds, a taped conversation between a general and a cosmonaut (occurring 20 years after the events of the drama) helps to explain aspects of the story.

The writer and production team envisaged content that could only be reproduced from the stereo bed, as well as broadly three types of audio content to be reproduced from the auxiliary loudspeakers.

1. *Speech and spot effects* (e.g., radio communication from base, specific sound effects). Replayed from appropriate auxiliary loudspeakers when available, or folded back into the stereo bed.
2. *Ambient sound effects* (e.g., weather sounds and atmosphere). Replayed from appropriate auxiliary loudspeakers when available, including duplication of the same audio to multiple devices. When no appropriate devices available, either folded into the stereo bed or removed from the mix as desired.
3. *Extra storyline* (the interview conversation between the general and the cosmonaut). Replayed from appropriate auxiliary loudspeakers when available, otherwise not heard at all.

Dialog and foley were recorded at Low Four, a music studio in the Old Granada Studios complex in Manchester, UK. Production for orchestrated devices requires an object-based format; however, in this case, standard radio drama recording techniques could be used, as there was no overlapping dialog and the majority of the sound design made use of effects libraries. Original music was also composed and delivered as multiple stereo stems.

3 Post-production

Prior to the production for orchestrated devices, a standard stereo version of the audio drama was created in *Pro Tools*. The recorded audio was combined with sound effects, either from libraries or generated using commercially-available plug-ins.

3.1 Orchestrated production setup

In order to convert the stereo production to a format suitable for reproduction with orchestrated devices, a bespoke object-based production environment was established. A system diagram is shown in Figure 1.

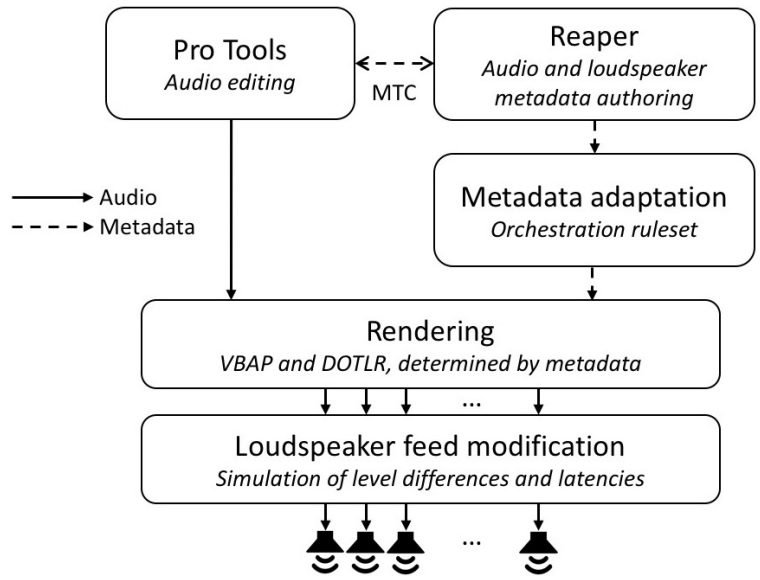

Figure 1: Orchestrated production system diagram

The sound designer continued to work in *Pro Tools* due to familiarity with the interface. Objects that were always present in the stereo bed were routed to a stereo bus, whilst objects that would either be part of the stereo bed or played from an auxiliary device—and those that would only play under certain conditions—were routed to additional mono buses. The buses were sent over MADI to the *Versatile Interactive Scene Renderer* (*VISR*) [5], which generated loudspeaker feeds (using amplitude panning (VBAP) for the stereo bed, and direct object-to-loudspeaker routing (DOTLR) for the auxiliary loudspeakers). Metadata (for stereo panning and the object routing rules detailed below, as well as to describe the set of loudspeakers) were generated by the sound designers using a set of bespoke plug-ins hosted in *Reaper* (Figure 2), and then sent to be processed in the *Metadapter* (a tool for adaptation of metadata based on a ruleset and user input [6]) before being passed to the *VISR*. The *Pro Tools* and *Reaper* sessions were synchronized using MIDI timecode (MTC).

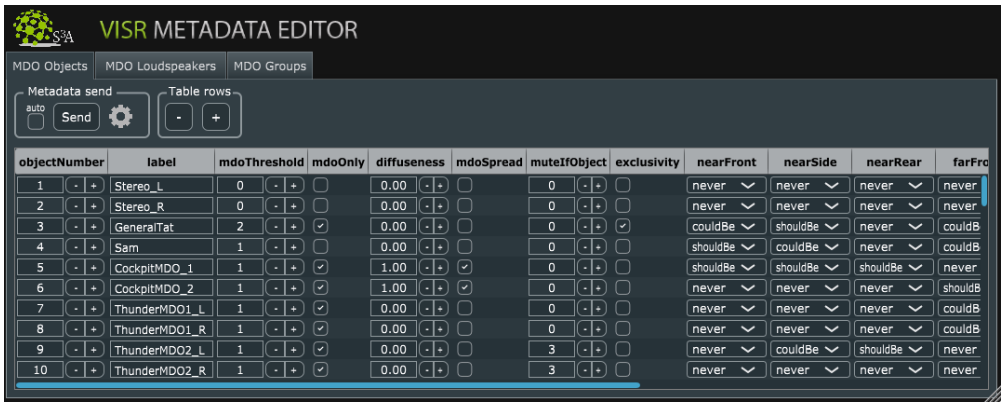

Figure 2: Metadata authoring plug-in hosted in Reaper. The production workflow also utilized a 3D panning plug-in.

The rendered content was replayed from a stereo pair of studio monitors (Genelec 8030B) augmented with a set of small consumer loudspeakers (Sony SRS-X11, termed "auxiliary loudspeakers" in this paper). All loudspeakers were connected using cables to ensure synchronization, and calibrated to produce equal sound pressure levels (SPLs) when placed at a fixed distance from the mixing position. The auxiliary loudspeakers were then placed around the mixing position in areas representative of the positional zones discussed below. A *Max/MSP* patch was included between the renderer and the loudspeakers to enable simulation of level, time, and frequency response differences between the loudspeakers. This was used to check the mix in conditions more similar to the target reproduction system. All aspects of the production environment were flexible so that the required metadata and rendering rules could be developed as new requirements were uncovered.

3.2 Metadata model and rendering rules

The metadata were authored in such a way as to enable flexible reproduction regardless of the available loudspeakers. The auxiliary loudspeakers were tagged with low resolution positional information indicating whether the speaker was near or far from the listener, and whether it was positioned to the front, back, or side of the listener.

For each object, metadata to facilitate the following rendering rules were captured (using the plug-in in Figure 2).

1. Zonal placement: for each positional zone, each object had metadata describing whether positioning the object in that zone was desirable ("should be"), permissible ("could be"), or precluded ("never").
2. Auxiliary only: objects could be precluded from the stereo bed (i.e., only played if a suitable auxiliary device was present).

3. Auxiliary device threshold: objects could be assigned to auxiliary devices only when a certain number of devices were available.
4. Device exclusivity: objects could preclude any further objects from being sent to the same device.
5. Selective reproduction: objects could be muted if other specified objects were being rendered (i.e., not absent after application of rules one and two).
6. Object duplication: after selection of suitable devices using the positioning information and above rules, objects could be replayed from all suitable loudspeakers (with or without decorrelation filtering) or a single loudspeaker (assigned randomly from the candidate set, prioritising "should be" zones over "could be" zones).

The rendering rules were implemented in a *Metadapter* processor [6], which applied the rules and adapted the metadata to instruct the renderer how to route the signals. Rules were applied to each object in order of ID number (objects with the 'device exclusivity' flag set were handled first).

4 Discussion

Producing audio for orchestrated devices is a large departure from standard production practices. It was consequently a challenging and time-consuming process. The following challenges were identified.

4.1 Writing and recording

The writer was initially unfamiliar with the orchestrated reproduction concept, and so found it difficult to visualize the final experience. Consequently, some promising ideas did not translate well to the mixing phase.

The difference in workflow from a standard production makes payment challenging—standard models may no longer be appropriate. Writers are often paid by word or by minute, but in this case the extra layers of content make these factors difficult to determine. Consequently, a lump sum payment was agreed.

For this piece, the recording process was similar to a standard radio drama. However, there were discussions about how best to record "extra storyline" content—as part of the full scene or separately. Ultimately, both were recorded; the separate takes were technically better and easier to mix, but the performance was better when the actors performed at the same time. Both types of take were used in the final edit. Actors initially responded negatively to the idea of "extra storyline" content that might mean their contributions were not heard. Clearer understanding of the technology might help to ameliorate these concerns.

4.2 Post-production

It was difficult for the sound designers to monitor the content given the range of potential reproduction layouts (variable number of devices, positions, qualities, output levels, and latencies). This was addressed by starting from a high-quality stereo mix, and augmenting it with a best-case MDO system (i.e., a high number of auxiliary loudspeakers in known positions with wired connections enabling synchronized reproduction at calibrated levels). Towards completion of the mix, the reproduction system was modified (e.g., by simulating delays or level changes and moving and/or removing loudspeakers). However, it is only possible to audition a limited number of combinations. The complex rendering rules required to orchestrate devices were difficult to understand and use, and continued to develop as requirements arose during the production. The low resolution positional categories and occasional random position assignments could be confusing for the sound designers, and it was difficult to break the habits of mixing to a known, fixed loudspeaker layout.

It was necessary to consider the likely limitations of the target loudspeakers (such as inaccurate synchronization, unknown but generally low output level, and poor bass response) when making mix decisions. For example, rhythmic musical elements requiring accurate synchronization were not routed to auxiliary devices. Having high quality loudspeakers as the main pair seems to give a large quality improvement, so it is beneficial to listen using lower quality main devices.

As some aspects of the content were only replayed when auxiliary devices were connected, it was hard to maintain an engaging narrative flow when no such devices were available. In the absence of the additional content, there were gaps in dialog that were longer than might normally be expected in standard audio drama production. This necessitated extra sound design, in some cases with objects that would only play when the additional content was not playing. A neater solution in future might be variable length content, which could be shortened if certain objects were not able to be played. Additional sound design elements were also required to make optimal use of the auxiliary devices and "fill out" the space. In some cases, this could be achieved by reproducing a sound from multiple loudspeakers, rather than adding extra sound design.

Some standard production techniques could not be used because of the flexibility of the reproduction. For example, sidechain compression is often used to "duck" the level of sound effects below speech. However, in the case of speech objects that weren't always present (due to the orchestration rules), ducking the sound effects in the absence of the speech would sound strange. In future, this could be addressed by having more object-based processing available.

The experimental nature of the production added complexity, compounded by differences in technical language used by researchers and sound designers. The metadata editor plug-in could be modified (by editing an XML configuration file) as needs arose, but did not have a particularly user-friendly interface. The *Metadapter* processor could also be expanded as rules were developed.

As a consequence of these challenges, the production took two weeks to mix—approximately ten times longer than a standard audio drama of the same length. Much of the first week was spent identifying challenges and developing technical solutions. It is likely that production time would be significantly reduced if another similar production was to be made, as much of the workflow has now been developed.

5 Summary and future work

A bespoke production environment was established to enable creation of an audio drama for orchestrated personal devices. Prototypes of similar experiences have been shown to be potentially beneficial; the drama described in this paper is the next step towards validating the concept and developing the production workflow.

Future work will focus on designing a public-facing delivery system, requiring: user-friendly onboarding and interaction (including device position reporting); accurate synchronization between devices; and implementation of the rules for object-to-device routing and rendering. This trial will be made publicly available on BBC Taster (http://www.bbc.co.uk/taster), a platform for delivering experimental productions to the public and collecting feedback.

The metadata model used for this production may not be universal; development of a flexible, reusable model is required. The rendering rules might also be extended; for example, further properties of the loudspeakers (maximum output level, frequency response, or latency) could be accounted for, and rules to handle devices being added or dropping out might be implemented. The content and metadata authoring tools should also be made easier to use. Further investigation into the appropriate use cases of the technology is also desirable.

6 Acknowledgements

This work was supported by the EPSRC Programme Grant S3A: Future Spatial Audio for an Immersive Listener Experience at Home (EP/L000539/1). *The Vostok-K Incident* was written

by Ed Sellek. Original music was composed by Oliver Vibrans. Production software support was generously provided by Giacomo Costantini and Andreas Franck. Details about the data underlying this work, along with the terms for data access, are available from http://dx.doi.org/10.17866/rd.salford.6984203.

References

[1] Jon Francombe et al. "Qualitative Evaluation of Media Device Orchestration for Immersive Spatial Audio Reproduction". In: *J. Audio Eng. Soc* 66.6 (2018), pp. 414–429. DOI: 10.17743/jaes.2018.0027.

[2] J. Woodcock et al. "A quantitative evaluation of media device orchestration for immersive spatial audio reproduction". In: *Audio Eng. Soc. Conf. on Spatial Reproduction*. Tokyo, Japan, Aug. 2018.

[3] M.O. van Deventer et al. "Media orchestration between streams and devices via new MPEG timed metadata". In: *International Broadcasting Convention*. Amsterdam, 2017.

[4] B. Taylor. "A history of the audience as a speaker array". In: *17th Int. Conf. on New Interfaces for Musical Expression*. 2017, pp. 481–486.

[5] A Franck and FM Fazi. "VISR — A Versatile Open Software Framework for Audio Signal Processing". In: *Audio Eng. Soc. Conf. on Spatial Reproduction*. Tokyo, Japan, Aug. 2018.

[6] James Woodcock, Jon Francombe, Andreas Franck, et al. "A framework for intelligent metadata adaptation in object-based audio". In: *Audio Eng. Soc. Conf. on Spatial Reproduction*. Tokyo, Japan, Aug. 2018.