Audio Engineering Society

Convention Paper 10203

Presented at the AES 157th Convention 2024
October 8–10, New York, NY, USA

This paper was peer-reviewed as a complete manuscript for presentation at this convention. This paper is available in the AES E-Library (http://www.aes.org/e-lib), all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Authoring Inter-Compatible Flexible Audio for Mass Personalization

Craig Cieciura[1], Elettra Bargiacchi[1], and Philip J B Jackson[1]

[1] *Centre for Vision, Speech, and Signal Processing, University of Surrey, Guildford, UK*

Correspondence should be addressed to Craig Cieciura (c.cieciura@surrey.ac.uk)

ABSTRACT

The popularity of internet-based services for delivering media, i.e., audio and video on demand, creates an opportunity to offer personalized media experiences to audiences. A significant proportion of users are experiencing reduced enjoyment, or facing inaccessibility, due to combinations of different impairments, languages, devices and connectivity, and preferences. Audio currently leads video in the maturity of object-based production and distribution systems; thus, we sought to leverage existing object-based audio tools and frameworks to explore creation and delivery of personalized versions. In this practice-based investigation of personalization affordances, an immersive children's fantasy radio drama was re-authored within five "dimensions of personalization", motivated by an analysis of under-served audiences, to enable us to develop an understanding of what and how to author personalized media. The dimensions were Duration, Language, Style, Device and Clarity. Our authoring approaches were designed to result in alternative versions that are inherently inter-compatible. The ability to combine them exploits the properties of object-based audio and creates the potential for mass personalization, even with a modest number of options within each of the dimensions. The result of this investigation is a structured set of adaptations, based around a common narrative. Future work will develop automation and smart network integration to trial the delivery of such personalized experiences at scale, and thereby validate the benefits that such forms of media personalization can bring.

1 Introduction

The popular conception of media personalization comes in the form of content recommendations as found in streaming services, for example. When we reflect, however, on the complexity of an individual's identity and attributes — as they pertain to experience and perception of media — including the contributory effect on quality of experience of permanent and temporary sensory or neurological impairments, environmental factors, device types and capabilities and the often transitory nature of ones' preferences, we soon realize that personalization can be a much broader concept.

In this study, we explored the re-versioning of a short radio drama excerpt as a vehicle to examine the multifarious ways it could be delivered each time for a different intended target user. We selected five categories, or dimensions, of personalization informed by an analysis of current broadcast provision which identified attributes of users that appear to be under-served, yielding an opportunity for the personalization function to serve them better. These dimensions are defined somewhat independently, which creates the possibil-

ity of combining them for mass personalization while retaining manageable levels of production effort.

Taking a practitioner-led approach positions this work between the capabilities of existing media distribution systems and state-of-the-art applied research, i.e., the limit of what is currently possible. Instead, using a collection of existing media production tools built around the concepts of object-based audio (OBA), both commercial professional packages and open-source tools/plugins, we sought to edit and re-mix the audio content manually for each of five dimensions of personalization whilst retaining inter-compatability between the techniques applied. This inter-compatability between the five dimensions, which themselves contain two, three or four versions, results in over a hundred potential combinations in which the radio drama can be personalized to a potential user. Building this understanding of how a practitioner would manually re-version audio content begins to provide a basis on which assistive tooling and automated procedures, likely incorporating machine learning techniques, might facilitate efficient provision of personalised media whilst maintaining narrative intent.

This paper first provides a short background on the subject of media personalization, primarily in the context of broadcast media (Section 2). Next, we highlight attributes of under-served audiences which informed our personalization dimensions (Section 3). We justify our requirements, in terms of content and selection of tools (Section 4), before detailing our personalization dimensions and the processes of re-versioning the content within each of them (Section 5). We describe how the different processes were brought together, enabling the inter-compatibility (Section 6). Finally, we summarise our work, recognize current limitations and present our next steps towards development of an end-to-end production and delivery pipeline (Section 7).

2 Background on personalization

During the early years of the media industry, as broadcasting was emerging a century ago, the primary drive was to establish distribution networks that could bring media content to people *en masse*, such as radio and talking pictures. Yet even then, the limitations of such a "one-size-fits-all" infrastructure were being questioned:

> "Radio is one sided when it should be two. It is purely an apparatus for distribution, for mere sharing out. So here is a positive suggestion: change this apparatus over from distribution to communication. The radio would be the finest possible communication apparatus in public life, a vast network of pipes. That is to say, it would be if it knew how to receive as well as transmit, how to let the listener speak as well as hear, how to bring [them] into a relationship instead of isolating [them]." Bertolt Brecht, *The Radio as an Apparatus of Communication* (1926).

Since then, the broadcast industry has evolved to encompass numerous methods of audio distribution and TV in all its forms. Yet, in the 21st century, audio streaming and video-on-demand have become mainstream, using a medium that supports two-way communication via internet protocol. Far from acting as a passive listener to the channel or station that the user has selected, the user's level of interactivity can extend to live streaming of user generated content and immersive online gaming.

Thus, this modern digital context for the large-scale distribution of media has potential to draw on a relationship with the user to shape the media that is supplied, either explicitly through direct interaction or implicitly from user profiles and interaction history. With this new opportunity to know more about the user and what they may want, the complementary technical challenge arises from the question: what things can we adapt in order to personalize the media service?

Prior works exist in the literature on creating personalized audio-visual media experiences. They include production of a one-off personalized cooking experience as a vehicle to begin development of an object-based media (OBM) workflow [1]; a description of a tool for building diverging narratives, *Storyformer* [2]; creation of two episodes of a medical drama with enhanced control over the audio mix via labelling audio objects within a hierarchical structure of narrative importance [3] and related prior work [4, 5, 6]. Other studies have examined these productions in the context of the experience of practitioners involved [7]; understanding user experience of object-based and personalized media [8]; and a review and assessment of production workflows in personalized media, resulting in propositions for the development required for personalized media offerings to become mainstream provision [9]. The last indicated a need for personalization capabilities to

be developed within the same ecosystem, or at least inter-compatibly, to afford each user the full breadth of options to personalize content to their particular combination of needs and preferences across media provision. This is as opposed to some personalization functionality being offered by one experience, other in another experience, and so forth with limited crossover, which has the potential to frustrate and alienate users.

3 Analysis of hard-to-reach audiences

To identify potential users who were better or less well served by current provision, we looked at statistics for the UK population and data relating to the UK's primary national broadcaster, the British Broadcasting Corporation (BBC). Viewing statistics identified sections of the audience who appeared to be well served, poorly served and hard to reach with current provision [10]. For example, 89% of audience in the ABC1 socio-economic group of 16+ adults reported use of BBC TV, iPlayer, radio or online on average per week from across the BBC's flagship services, and 97% of the over 55 age category. However, in the 16 to 34 age bracket only 76% were weekly audience and only 73% of the Black, Asian and Minority Ethnic adults. These differences between audience segments suggest some level of inequity in current provision and indicate a potential opportunity for personalization to address this.

Language. Considering the BBC's UK reach by language, the three main native languages cover 92% of the population's primary language: English (60M), Scots (1.6M) and Welsh (1.1M) from a total of 67M. There are 20 languages with populations >100,000 in the UK, accounting for 5% of population, and 80 further languages needed to reach the remaining 3%.

The reach of these languages worldwide paints a different picture. English serves just 19% of the global population, combining first and second language speakers. There are 15 languages with populations over 100M. The top 20 languages cover 81% of the world's 8bn population. The BBC's international services extend to 43 languages and reach around 450M each week. From this perspective, any serious attempt to reach a global audience must consider alternative language versions.

Human conditions. Sensory impairments affect a significant fraction of audiences. In the UK, $\sim$12M are hearing impaired (18%) including 87,000 members of the deaf community and $\sim$2M are visually impaired (3% population) with 340,000 blind/partially sighted. As dialogue is often central to audiences' enjoyment of the content's narrative, problems with intelligibility have for years been a major cause for complaints to broadcasters. To widen accessibility, a proportion of the BBC's output, for example, is given subtitles (88%), sign interpreting (7%) and audio description (19%).

Cognitive variations in mental function, often relating to attention and memory, including ADHD, dyscalculia, dyslexia, dyspraxia and autism, form a set of conditions associated with neurodiversity. These are estimated to affect around 10M (15%) in the UK. Mass personlization provides an opportunity to make the full breadth of content more accessible for users with sensory or cognitive differences, which can be complex, interlinked, and interact with other factors (such as device, environment and social setting).

Devices and Connectivity. For some users, limitations of their play-out conditions constrain their access to content, including media devices with low-quality screens or loudspeakers, extreme lighting, loud background noise or reverberation, and poor network connectivity. In terms of TV broadcast in the UK, around 1.5% of premises are without coverage; similarly, 1.3% of homes are without decent broadband, whereas 7% are without decent mobile network coverage [11]. On average, around 10% of connections are poor, in rural locations, and often interrupted when travelling. When considering how users are accessing digital videos, 63% use Smart TVs and only two percent fewer, 61%, use Smart Phones. Those using the latter are significantly more exposed to issues related to limited or variable connectivity, environmental factors, such as excessive background noise which can be intrusive with or without headphones, and hardware limitations.

A 2018 study indicated that only 13.5% of UK households own an audio “surround system" [12]. Since then, the concept of immersive audio has been somewhat reinvigorated in the public discourse, e.g., Apple Spatial Audio, Dolby Atmos, etc. Despite this, immersive audio content is of comparatively limited availability. Users who are seeking to enhance their experience are also considered in our authoring of personlization.

Age. A final perspective that we have considered is age, which has a strong interaction with the kinds of media platform that dominate users' viewing habits. In general, younger audiences lean comparatively

more towards user-generated content and social media, whereas older audiences are more likely to rely on traditional broadcast provision. As such, The BBC's on platform weekly reach is 97% with those aged 55 and over, but only 76% for those aged 16 to 34, suggesting that young people are up to eight times more likely to miss broadcaster-generated content [10, p. 12]. We are interested to explore therefore whether targeted adjustments of the content to appeal to this younger audience, almost a quarter of the population [13], can achieve a more even balance.

The factors that we have discussed above are not exhaustive. Most geographic, cultural and demographic factors, including protected characteristics under the Equality Act 2010, show some imbalance and other significant under-served groups may be identified. Our main aim has been to identify the largest groups where personalization has potential to offer a transformative gain, such as enabling access. Yet it is important to acknowledge that the factors assume one context in the UK, they are provisional and changing, although they are general enough to allow potential translation to many other contexts of media distribution elsewhere.

3.1 From audience needs to personalization dimensions

Based on this analysis of user needs, we chose to focus on five dimensions of personalization that independently change the standard mix, i.e., to alter:

Duration: for users who require a different pace whilst not substantially altering the narrative arc or key moments — the narrative intent.

Language: presenting the narrative structure across different languages, which may have different lexical densities, whilst maintaining chosen duration.

Style: adapting to different preferences or age ranges whilst maintaining the core narrative structure and maximizing re-use of audio objects.

Device: adapting the content to give an optimal experience for the capabilities of the device, from low-quality mono up to high-quality immersive.

Clarity: to improve comprehension of the narrative for users for whom it is reduced by sensory or neurological conditions, poor reproduction capabilities or factors in the environment.

Whilst seeking to ameliorate the challenges of users who fall into the categories of *hard to reach audiences*, an additional motivation is the observation that many technology features that have been designed with a specific need in mind have become mainstream features with a wide user base, improving usability across a range of contexts. Classic examples of this include the modern uses of subtitles by those with normal hearing and the recent popularity of podcasts and audiobooks.

Furthermore, the important recognition of the complex and multi-factorial needs of the modern audience inform our aim to retain inter-compatability between the dimensions. This in turn leads to multiplication of our different versions from one original, to alternatives within five dimensions, to hundreds of possible combinations of affordances.

4 Planning and preparation

To select a piece of content for our study we derived a set of requirements. First, that original production used an object-based workflow, i.e., consisting of chunks of audio and associated metadata describing the their attributes, to enable flexible reworking. Next, for the original experience to be dramatic and fictional, to provide the greatest scope for adaptation whilst also providing constraints that the content was narrative driven and therefore not trivially segmentable.

Development of object-based audio, as opposed to video, has seen comparatively greater advancement in, for example, development of object-based frameworks, such as [14], published as standards, e.g., [15, 16]. This development has seen high quality example content produced and distributed in OBA formats. One such example is the S3A Object-Based Audio Drama Dataset [17], which consists of three OBA radio drama scenes. We selected *The Forest Scene*, 04:39 in duration, described as a fantasy scenario set in a forest. Opening with two children playing in an autumnal forest, after one runs away the other encounters a large, friendly monster. After falling into a river, the child is rescued by the monster and experiences a journey through a landscape experiencing changing seasons. The experience contains a narrator, non-diegetic music and a multitude of sound effects.

A later release of *The Forest Scene* is provided in the context of a suite of object-based production tools [18]. We utilised this release but moved the assets to a new session set up using the open source EAR Production Suite[1] (EPS) [19], which superseded the tools

[1] https://ear-production-suite.ebu.io/

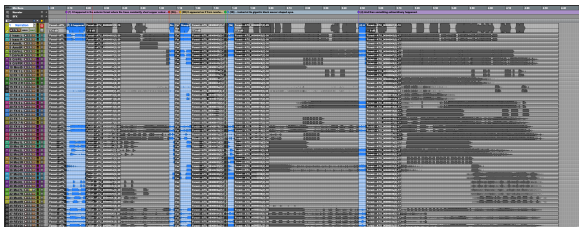


Fig. 1: Pro Tools[2] project with (in blue) key dialogue, associated music and sound effects.

described in Costantini et al. [18]. The EPS is a VST plugin suite for authoring codec-agnostic spatial audio productions in the standardised ADM format [14]. Due to familiarity, Avid Pro Tools and the Dolby Atmos Production Tools were also used for composition and editing sub-tasks and for later stages of mixing.

Several preparatory tasks were performed consisting of processing each of the audio objects to remove noise, editing out silence, and applying basic equalisation.

5 Dimensions of personalization

Exploration of the re-versioning of the audio content was performed via a collaborative and iterative process, consisting of editing/remixing then discussion and feedback, repeated to refine each of the versions until they were considered complete. After completing each version the practitioner conducted a reflection task, writing up a series of notes describing challenges, decisions and techniques. In the following sub-sections this reflection is distilled into a description of the process of producing each version of the experience.

5.1 Duration

First, we explored altering the *duration*, or pace, of the experience. For a faster or slower paced story, a shorter or longer duration alters the form, while seeking to maintain the narrative intent. Both alterations have their use. Here, to obtain an illustrative test example, we explored a 75% reduction, from the original 04:39 duration to 01:10, and then again by another 50% to 00:35. A workflow for abbreviating was developed, which can shorten more or less. First, the script, obtained through dialogue transcription, was analysed to identify key elements of the narrative structure:

- protagonist voice: narrator embodies the child;
- antagonist appears: monster approaches;
- first incident [20]: meeting of child/monster;
- second incident: child encounters the water;
- common endpoint: monster and child travel to the winter forest.

Table 1: Excerpt of script changes, deleted dialogue in italics: (a) scene introduced, (b) description removed, (c) environment illustrated in sound design, (d) characters' interaction, (e) key description abridged.

It happened in the Autumn forest, where the trees constantly shed copper coloured leaves and there's always chill in the air.[a]
Amelia and I loved it there.[b]
It was heaven for our eight year old selves.
Miles and miles of space to run around trees so tall and wide that we could chase each other and never be found.[c]
When we were there, we became Nathan and Amelia, the Prince and Princess of the Autumn Forest, except one day when Amelia ran off without me.
That was the first time I saw it. It appeared as if from nowhere, staring at me.[d]
I *faced it,* looked at its gigantic black saucer shaped eyes…[e]

An excerpt, in Table 1, displays dialogue from original and short versions, illustrating identification of the key moments and the lines of dialogue critical for the narrative intent. Lines containing redundant information were compressed, e.g., "I [*faced it,*] looked at its giant black saucer shaped eyes". Figure 1 illustrates the audio editing following this script analysis. Next, the original version's music and sound effects were examined, including ambient sounds in the scene, e.g., birds and crickets, and action sounds, such as the creature's heavy footsteps. These were not always locked in alignment to the dialogue. For compactness, the timing of key sound effects was adjusted where phrases were removed for the short version, such as children running, playing and laughing. The practitioner consistently employed this approach, identifying and re-using sound effects to represent meaning from the original dialogue and its narrative intent. In places, sound effects were repeated, such as chimes or roars; elsewhere new sounds were integrated.

A similar approach was adopted for the music, which is continuous, but using an alternative editing method. Rather than plucking out discrete sound effects, the music aligned with temporal adjustment to the original

narrative moments was blended to provide seamless transitions between each narrative section. Dialogue timing was paced to flow naturally, in relation to key sound effects and musical changes, avoiding clashes.

5.2 Language

Offering alternative language access to media experiences as a mainstream personalization option differs from dubbing of films or TV series. Ideally, users would expect consistent personalization options to be available across content for a wide range of languages. In the future, machine-learning based approaches might be utilised, the interaction between the pacing of the narrative and other audio components was examined via this manual intervention.

The second author and main sound designer is fluent in Italian, German and English. Considering these languages initially, a language translation workflow, intended to be inter-compatible with duration changes, was developed. Seeking to test this process for unknown languages, for future automation, traditional Chinese (i.e., Cantonese) was added.

Initially, a full word-for-word translation from English to German was performed and recorded. Yet this did not allow intra-sentence editing, as described in Section 5.1 to adjust duration. In German, the word sequence construction changes between main and subordinate clauses, restricting duration editing on the word-for-word German translation. For example, when "and that's when I realized, I was in the winter forest" was shortened to: "I was in the winter forest", the word-for-word translation failed grammatically.

Another approach was tested, dividing the dialogue into atomic phrases that needed no further editing and could be recombined to create varying durations which matched across languages —– the latter was challenging when the full translation was performed. Additionally, sentences requiring intra-sentence editing when altered for duration were identified. Together, both sets of translated sentences provided duration flexibility.

In the transcribed dialogue, punctuation was interpreted and reflected the original English performance. In language translation, punctuation choices affect phrases' duration, providing some control, e.g., the comma in "Because I was so high and the ground was so far away, but the thing didn't seem to mind" provided a better match in Italian, which would otherwise have made two

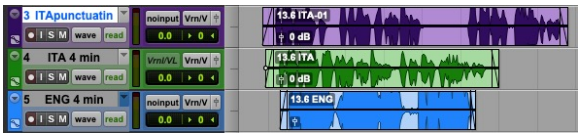


Fig. 2: A sentence in original English (blue) and Italian with a period (purple) and comma (green) which approximates the required duration.



Fig. 3: Manual adjustment of center-aligned clips to prevent overlap with key music and sound effects, prior positions shaded yellow on Italian (green) and German (orange) tracks.

sentences and overrun. Figure 2 displays the duration of the clips in the DAW.

A process for aligning atomic dialogue chunks with the other audio components in re-timed alternative language versions was developed. The music and sound effects appear before the first dialogue, so alternative versions were aligned to the original (English) onset. Conversely, for the final clip, versions were aligned ending before the outro music and sound effects. Intermediate clips were center aligned by default, with some manual adjustment to prevent longer translations crashing key sound effects or music.

When dialogue was denser, longer alternative-language chunks sometimes overlapped when following this simple method. When determining how to proceed, the boundaries of narrative segments and issues of overlapping and masking key sound effect or music elements had to be considered. The narrative segment boundaries and key events prescribed hard boundaries for the dialogue. A naturalistic result was obtained by manually refining timing adjustments within these bounds.

5.3 Style

In the style dimension, the power of music and sound effects to dramatically alter the mood or feel was utilised.

From a creative perspective, there is a tension between the director's vision and the listener's individual preferences. Superficial changes can make content more attractive for certain audiences, whereas more substantial changes challenge the narrative's integrity. Here we explored a "horror" stylization of the original fantasy story to address a slightly older teenage audience, in contrast to the original primary school age target of the original. To enable interchange between multiple personalizations in combination, we maintained consistent structure of our different versions, anchored by key narrative moments and distinct audio layers for the narrator (which remained unchanged), actors, music and sound effects in the object-based mix.

In the original version, only mild peril is implied with very brief moments of tension which are quickly resolved. The monster was portrayed with predominantly friendly or neutral sounds, after the initial roar. The ending was suspended, with a curiosity-inducing but not negative conclusion. The horror version offers a darker, scarier mix targeted for teens, striking a contrast against the benign original that was aimed at children. It emphasised negative connotations with the forest, while the original was generally positive. A scary atmosphere was created with suspense, rising tones, audible heart beats, dramatic music and a shocking end.

The choice of music was fundamental to establishing the style's mood, to create the emotional frame and atmosphere of the story, tension or joy. The sound effects reinforce this. Naturally, the mix also has an effect. The Horror version is more prominent in the low frequencies (LF) and has a stronger use of reverb, and is generally more aggressive. Whilst aiming to match the same loudness (-23 LUFS), these choices resulted in a mix approximately two LUFS louder.

As the scene is set in the first four seconds, sound effects were selected to suit the desired temperament of the horror version.There are wind, water and birds: strong and threatening wind versus a light breeze; a frightful thunder storm versus a gently babbling stream; crows cawing versus chirpy robins. You can hear children's voices, breaths and actions: worried breaths versus happy laughter.

The character of the monster was altered through sound. The monstrous features, already present in the original, are expanded in the Horror version: more frequent and stronger roars, heavier footsteps by intensifying the LF. In the original version, the creature is a giant and of initially questionable intention, but quickly revealed as friendly. Its steps indicate huge size, but are not ominous. The first roar is mildly scary, but quickly turns to non-threatening grunts and vocalizations.

In the ending, these two versions diverged with drastically different story developments, even though the narration remained exactly the same: words, voice and intonation. In the horror version, the child and monster come into direct conflict and the child is cast away with a terrifying atmosphere, and horrific scream. In the original version, the sound effects suggest a more magical mystery, but the child reacts full of wonder, fun and curiosity. Thus, not touching the narration, the transformation between these two versions was achieved entirely via music and sound effects.

The object-based nature of the production offers the potential to reuse source audio between different versions, where appropriate, with differing parameters encoded as metadata. Whilst not the subject of this paper, maximising audio reuse through applying, to the same clip, varying volume automation, panning and processing dependent on the intended version would minimize the data required to be stored and sent to users.

5.4 Device

The clearest distinction between reproduction formats is that of stereo versus immersive —– surrounding in the horizontal and, at least somewhat, vertical planes. In the UK, the largest broadcasters primarily offer their audio component in stereo, including via their VOD services. Other online distributors request delivery from content producers in stereo and some variation of surround sound or immersive formats, such as Dolby Atmos. This distinction is present in loudspeaker-based reproduction with many users primarily utilising the loudspeakers built in to their televisions [12] and a smaller segment surround-sound. For headphone-based listening, stereo is most typical with a limited binaural offering.

A key benefit of immersive reproduction, with physical loudspeakers or binaural, is the ability, logically, to enhance immersion, ranked the most important perceptual attribute in spatial audio replay [21]. Beginning with a typical stereo mix for exploring the dimensions previously described, we then sought to create immersive versions with a 7.1.4 setup, emphasising the additional potential of this format whilst retaining inter-compatibility between the other dimensions.

Narrator. The narrator is this piece's most important element, panned centrally in the stereo mix. In the immersive version, the absence of a visual reference was exploited to free this sound object. The narrator's voice was spread across all the main loudspeakers, enveloping the listener from all sides. Perceptually, the result is a voice's presence but without identifiable location, magnifying the fantastical nature of the piece.

Sound effects. Given the setting of the original piece, a fantasy forest full of creatures, the immersive mix offers an amplification of detail and spatialisation. The mix includes an LFE channel with dedicated audio content, and multi-channel environments (e.g., distant wolves and ambient crickets as 5.0 beds). To enrich the soundscape texture, the ambience tracks were layered and re-positioned to distribute sounds all around: for the wind, two objects moved independently at different elevations, to give a more natural impression of movement with gusts of wind. The leaves rustling in the breeze are positioned in the upper layer to evoke the height of the trees. Similarly, the upper layer adds to the monster's huge scale, from the child protagonist's point of view. Fluidity and localisation of the movement is enhanced, e.g., the children running around the forest, with their steps, voices and brushes against the foliage, are positioned in the more dense middle layer. These choices demonstrate that a contextual linking of the relationship of sound objects to the narrative is key and determination and encoding of these features would be necessary in an (semi-)automated system.

Music. With most of the action at ear height, the upper layer provides space for the music. Interesting passages were embellished with movement, e.g., by throwing the arpeggio at 0:30 back and forth, following its ascending-descending pattern.

Summary. The immersive version locates the listener inside the story utilizing the middle layer for narratively important sound effects, with a focus on spatialization and detail of the environment, and the upper layer for music and additional scene-setting ambience. The narrator was distributed across all channels, giving an unreal quality and clearly presenting its different function from the other sounds. In contrast, the stereo version is, by nature, simpler spatially. The mono version necessitates a reduction in density of sound effects to avoid masking key components of the mix. Considering the challenge of automation, an approach of first identifying and emphasising these narratively important elements, as was the manual method here, offers an initial direction. This could be followed by the use of more efficient tools to insert sound elements, for upmixing from stereo, and generative AI.

5.5 Clarity

Clarity is a semantically overloaded term within the audio field. In this context we use it to refer to a version created with the desire to enhance the comprehension of the narrative via improving the intelligibility of the dialogue, foremost, and of sound effects that sign-post the narrative. The mix was approached with the broad intention of being a beneficial version in various situations: e.g., a person who suffers from hearing impairment, or another listening in noisy or reverberant environments, or with a low-power or low-quality system, e.g., a small Bluetooth loudspeaker or an old mobile phone. The issue of intelligibility is a well-documented concern, particularly for broadcasters, as it cuts to the heart of whether a story told is actually heard. Good intelligibility is therefore an essential pre-requisite for effective delivery and enjoyment of the narrative. Yet in certain stages of our lives and many situations in our daily lives, our access for understanding the audio content is restricted. Prioritising the sound's intelligibility within the clarity version aims to alleviate this restriction.

It is widely recognized that aging is frequently associated with a decline of hearing sensitivity (presbycusis) [22]. With changing demographics, the proportion of audiences with age-related hearing loss is likely to increase, a large and growing population [13]. Hearing loss is typically characterised by a gradually declining sensitivity at higher frequencies. For speech at normal levels, such listeners lack a significant portion of speech cues. In addition to hearing loss, older adults tend to lose temporal auditory processing abilities, meaning they are less able to detect brief gaps in sound. Interference from background noise impairs speech comprehension, which compounds with hearing loss.

These concepts guided us in the plans and choices we made in our preliminary aim at a clarity mix, recognizing the vast existing literature, including [23, 24, 4, 5, 6], and the need for subsequent testing. We first simplified the sound material: reduced the number of sound components, eliminating sounds less relevant to the narrative, to reduce interference and distraction from the key elements, such as the narrator and characters.

Next, LFE was removed and the spectrum of the material was band-limited between 70 Hz and 18 kHz to make reproduction more timbrally consistent. Reverb was reduced or removed on foreground objects. Space, amongst the sound effects and music, was created to spotlight the key elements in the mix. Additional EQ boosts were added to the already processed dialogue (+2 dB at 1 kHz and 4.7 kHz), a frequency where the gap between speech information and hearing loss is particularly wide [22, 23]) plus a multi-band dynamic EQ and compressor. Dynamic EQ was used on sound effects to make space, spectrally, when the dialogue is present. Music was reduced in relative level. Atmospheric sounds were reduced in complexity and used more sparingly. Volume automation was employed to carve out additional space for dialogue and narratively important sounds.

Simulating adverse conditions, the clarity mix was auditioned via reproduction on a small Bluetooth speaker in the presence of multi-channel noise that was rendered from third-order Ambisonic recordings, and additional sound design, of kitchen and aircraft cabin environments. The literature on hearing impairment justifies further work to refine the approach to address other deficits, e.g., in temporal auditory processing. Formal listening tests may validate efficacy and user benefits.

6 Structuring for inter-compatibility

After developing authoring processes for each of the dimensions somewhat independently, the next stage was to bring these processes, and the audio assets, together into one session. The combined session, nicknamed *macro-session*, enabled switching, combining and auditioning of and between each dimension and version. This was achieved by importing all of the audio assets whilst maintaining their relative temporal positions and loudness. The session was structured in the form *source -> object -> scene*, an additional layer to the typical *object -> scene* structure within an object-based production. This structure was used to prevent duplication of audio assets within the project, and therefore within a later object-based export (such as within an ADM-based delivery file) by separating different processing applied to the same source audio, for different versions, to the secondary *object* stage. Sources, e.g., the monster footsteps in the original and horror style versions, have alternative volume automations, panning, EQ and reverb applied at the *object* stage, without duplication of the audio itself. The ADM standard [14] includes parameters for gain and panning. If audio processing (e.g., EQ and reverb) were to be offloaded to edge compute or user devices to be applied on-demand, reducing the need to store and transmit multiple versions of audio data based on the same source, additional metadata would be required to encode these parameters. The final stages of editing and mixing required careful timing, level and spatial alignment of assets in the macro-session, which enabled rapid previewing of combinations to ensure inter-compatibility across all $2\times3\times2\times5\times2=120$ versions with personalization.

7 Summary and future work

Using a short children's radio drama piece, we presented a practice-based study of personalization around five inter-compatible dimensions of variation: duration, language, style, device and clarity. These dimensions were motivated by under-served audience categories. We described the main considerations and techniques applied in the adaptation of each of the versions, the challenges faced and the development of generic approaches. The challenge of mass personalization is the consistent provision of affordances for the complex and interacting needs of audience members which requires many alternative versions. Maintaining inter-compatibility throughout results in the ability to combine the affordances of different versions resulting in 120 combinations from our manually authored dimensions. This vastly enhances the potential to improve experiences for the diverse needs of audience members.

To enable consistent provision, the impracticality of manually authoring multiple alternative versions of any piece of content must be recognised. In our future work, by modelling the manual processes described in this paper, we will seek to partially-automate the authoring process whilst cognisant of the need to maintain production intent and will trial and validate this approach on other use cases. By taking a rigorously object-based approach, we aim to minimize duplication of media in storage and transmission by further separating the source media, metadata and processing, the latter where appropriate, taking advantage of IP-based delivery and the increased processing power of modern media reproduction devices. This will enable content producers to offer a vast array of versions taking advantage of object-based media's flexibility.

8 Acknowledgements

This research was funded by EPSRC-BBC Prosperity Partnership 'AI4ME: Future personalised object-based media experiences delivered at scale anywhere' (EP/V038087/1). For the purpose of open access, the authors have applied a Creative Commons Attribution (CC BY) license to any Author Accepted Manuscript version arising.

References

[1] Cox, J., Brooks, M., Forrester, I., and Armstrong, M., "Moving Object-Based Media Production from One-Off Examples to Scalable Workflows," *SMPTE Motion Imaging Journal*, 127(4), pp. 32–37, 2018, ISSN 2160-2492, doi:10.5594/JMI.2018.2806499.

[2] Armstrong, M., Bowman, S., Brooks, M., Brown, A., Carter, J., Jones, A., Leonard, M., and Preece, T., "Taking Object-Based Media from the Research Environment Into Mainstream Production," *SMPTE Motion Imaging Journal*, 129(5), pp. 30–38, 2020, ISSN 2160-2492, doi:10.5594/JMI.2020.2990255.

[3] Ward, L., Paradis, M., Shirley, B., Russon, L., Moore, R., and Davies, R., "Casualty Accessible and Enhanced (A&E) Audio: Trialling Object-Based Accessible TV Audio," in *The 147th Audio Engineering Society Convention*, New York, NY, USA, 2019.

[4] Ward, L., Shirley, B., and Francombe, J., "Accessible Object-Based Audio Using Hierarchical Narrative Importance Metadata," in *The 145th Audio Engineering Society Convention*, New York, NY, USA, 2018.

[5] Ward, L. A. and Shirley, B. G., "Personalization in Object-based Audio for Accessibility: A Review of Advancements for Hearing Impaired Listeners," *The Journal of the Audio Engineering Society*, 67(7/8), pp. 584–597, 2019.

[6] Shirley, B. and Ward, L., "Intelligibility versus Comprehension: Understanding Quality of Accessible next-Generation Audio Broadcast," *Universal Access in the Information Society*, 2020, doi:10.1007/s10209-020-00741-8.

[7] Ward, L., Glancy, M., Bowman, S., and Armstrong, M., "The Impact of New Forms of Media on Production Tools and Practices," in *IBC2020*, 2020.

[8] Maxine Glancy, Lauren Ward, Nick Hanson, Andy Brown, and Michael Armstrong, "Object-Based Media: An Overview Of The User Experience," in *BBC White Paper*, volume 390, 2020.

[9] Cieciura, C., Glancy, M., and Jackson, P. J., "Producing Personalised Object-Based Audio-Visual Experiences: An Ethnographic Study," in *Proceedings of the 2023 ACM International Conference on Interactive Media Experiences*, IMX '23, pp. 71–82, Association for Computing Machinery, New York, NY, USA, 2023, ISBN 9798400700286, doi:10.1145/3573381.3596156.

[10] BBC, "BBC Group Annual Report and Accounts," Technical report, 2023.

[11] "Connected Nations Update: Summer 2023," Technical report, Ofcom, 2023.

[12] Cieciura, C., Mason, R., Coleman, P., and Paradis, M., "Survey of Media Device Ownership, Media Service Usage, and Group Media Consumption in UK Households," in *The 145th Convention of The Audio Engineering Society*, AES, 2018.

[13] James Robards, "National Population Projections - Office for National Statistics," Statistical Bulletin, Office for National Statistics, 2022.

[14] "ITU-R BS.2076-2: Audio Definition Model," 2019.

[15] "ISO/IEC 23008-3:2022: Information Technology — High Efficiency Coding and Media Delivery in Heterogeneous Environments — Part 3: 3D Audio," 2022.

[16] "ETSI TS 103 190-2 Digital Audio Compression (AC-4); Part 2: Immersive and Personalized Audio," 2018.

[17] Woodcock, J., Pike, C., Melchior, F., Coleman, P., Franck, A., and Hilton, A., "Presenting the S3A Object-Based Audio Drama Dataset," in *The 140th Audio Engineering Society Convention*, 2016.

[18] Costantini, G., Franck, A., Pike, C., Francombe, J., Woodcock, J., Hughes, R. J., Coleman, P., Whitmore, E., and Fazi, F. M., "A Dataset of High-Quality Object-Based Productions," in *The 147th Audio Engineering Society Convention*, Audio Engineering Society, 2019.

[19] M Firth, R Bailey, and C Pike, "The EAR Production Suite: Next-Generation Audio Software Tools," 2020.

[20] Field, S., *Screenplay: The Foundations of Screenwriting*, Delta, 2005.

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

[22] "BS EN ISO 7029:2017: Acoustics. Statistical Distribution of Hearing Thresholds Related to Age and Gender," 2017.

[23] Müsch, H., "Aging and Sound Perception: Desirable Characteristics of Entertainment Audio for the Elderly," in *The 125th Convention of the Audio Engineering Society*, Audio Engineering Society, 2008.

[24] Armstrong, M., "From Clean Audio to Object Based Broadcasting - BBC R&D," White Paper 324, BBC Research and Development, 2016.