OPEN ACCESS Freely available online

Journal of the Audio Engineering Society
Vol. 65, No. 4, April 2017 (© 2017)
DOI: https://doi.org/10.17743/jaes.2017.0005

Personalized Object-Based Audio for Hearing Impaired TV Viewers

BEN SHIRLEY, *AES Member*, **MELISSA MEADOWS, FADI MALAK,**
(b.g.shirley@salford.ac.uk)

JAMES WOODCOCK, AND ASH TIDBALL

Age demographics have led to an increase in the proportion of the population suffering from some form of hearing loss. The introduction of object-based audio to television broadcast has the potential to improve the viewing experience for millions of hearing impaired people. Personalization of object-based audio can assist in overcoming difficulties in understanding speech and understanding the narrative of broadcast media. The research presented here documents a Multi-Dimensional Audio (MDA) implementation of *object-based clean audio* to present independent object streams based on object category elicitation. Evaluations were carried out with hearing impaired people and participants were able to personalize audio levels independently for four object-categories using an on-screen menu: speech, music, background effects, and foreground effects related to on-screen events. Results show considerable preference variation across subjects but indicate that expanding object-category personalization beyond a binary speech/non-speech categorization can substantially improve the viewing experience for some hearing impaired people.

0 INTRODUCTION

Many older people have age related hearing loss that is detrimental to the viewing experience of broadcast media. Over recent years some considerable effort has gone into investigating the issues that people with hearing impairments face in their experience of television broadcast. Research and surveys by the BBC, research sponsored by the Independent Television Commission and Ofcom in the UK, and also by the European Commission means that these issues are understood much more clearly than was once the case. However there still remain important outstanding perceptual and technological research questions that need to be answered in order to most effectively deliver better audio for hearing impaired people. This paper presents the development of an object-based audio approach to providing clean audio for hearing impaired people using an open object-based audio format. Object-based audio differs from current, channel-based audio in that instead of being defined with reference to a loudspeaker layout (2.0, 5.1, etc.), elements of a sound scene remain as separate objects that are defined by metadata. Metadata can include such information as, for example, the coordinate location of a sound source, its level, and directional characteristics. Object-based audio has largely been associated with immersive 3D audio systems but can also facilitate personalization, interaction [1], and can be reproduction system agnostic [2].

Perceptions of several genres of broadcast media sound scenes were investigated using an elicitation approach to identify perceptually distinct audio object categories. These categories were used to create an object-based audio mix, which was presented to the hearing impaired participants using an interactive user interface, allowing participants to set their own preferred levels for each category stream. Object-based audio personalization based on audio object categories showed real benefits for hearing impaired people that can be realized as soon as object-based programming becomes mainstream and is broadcast to viewers.

1 CLEAN AUDIO FOR TV BROADCAST

Issues associated with the experience of television viewing for people with hearing impairments have been much discussed over recent years. The ITU recognized what it called an "increasing incidence and awareness in the community of hearing impairment" and raised a question as to what an appropriate relationship between dialogue, music, and effects may be for hearing impaired viewers [3]. The ITU report also questioned how these parameters might be determined, how such audio media may be most effectively

produced, distributed and transmitted, and what facilities might be provided at the receiver. These areas have formed the basis of much of the research into clean audio for TV broadcast since.

1.1 Research into Hearing Impairment and Broadcast

The BBC's position as a public service broadcaster in the UK, together with their substantial research base, has placed them at the forefront of accessibility research relating to broadcast and related media. The BBC's early work into TV audio for hearing impaired people [4] was carried out in response to complaints from viewers about background sound, such as audience laughter, crowd noise, and background music that made speech difficult to understand. The work recognized that excessive levels of background sound would impair the enjoyment of viewers, both with and without hearing impairments, and looked at perceived intelligibility of dialogue in test material comprising a number of TV programs with varying levels of background sound. Results were inconclusive; the test was quite large scale with 336 test subjects but, inevitably in such circumstances perhaps, took place in largely uncontrolled listening environments that may have had a confounding effect. Indeed, Mathers identifies other confounding factors common to most research carried out in this area: tests are usually carried out with sound with accompanying video therefore making some degree of lip reading likely, and the self-selecting nature of participants in this type of research makes any panel unlikely to be representative of any populations' hearing abilities. Indeed, the wide variation and individuality of hearing impairments in any population presents its own challenges when looking for a "one size fits all" solution for clean audio. Further challenges to the assessment of clean audio solutions are highlighted by Carmichael et al. [5] in the VISTA project where a high degree of "speech reading" was recognized as being attempted by older participants in attempting to understand an avatar with a synthetic voice. This was partially unsuccessful owing to lip sync problems although this in itself indicates a degree of reliance on visual cues for older users. Other research [6] shows biasing in assessments of AV media quality from both audio and visual interactions. The research indicates that, in their study, quality of visual presentation had more impact on assessments of audio quality than quality of audio presentation had on assessments of visual quality. In each case significant influence was demonstrated. For the audio researcher this is potentially problematic and care must be taken to ensure that potential biases are minimized.

In other BBC research discussing the potential benefits of surround sound broadcast Meares [7] discusses the potential to broadcast a separate hearing impaired (HI) channel using the same method as is currently used for alternate language. This approach relies on having access to separate components of the mix at the production stage in order to deal with dialogue separately. For media where access to original unmixed audio stems is available broadcasting a separate HI channel is an attractive solution, however discussions with broadcast industry research partners indicated an unwillingness to pay for the additional bandwidth required.

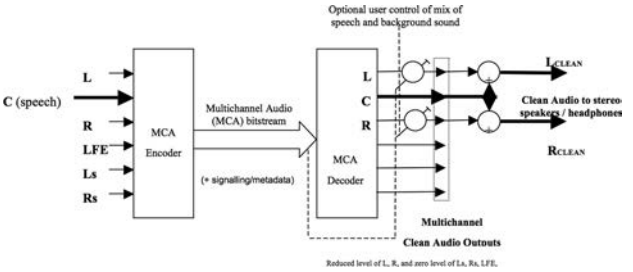

Fig. 1. Example Clean Audio system reference model based on speech in center channel, reproduced from [12].

Other work on improving TV sound for people with hearing impairments has focused on speech enhancement; separation of speech from competing audio sources at the set top box, an area of research that has challenges in common with the blind source separation (BSS) problem [8]. A summary of these approaches is presented by Armstrong [9] who concludes that, "audio processing cannot be used to create a viable 'clean audio' version for a television audience." The factor differentiating speech enhancement for television broadcast from pure BSS problems however is that broadcast media generally conforms to known production conventions and standards and this can be leveraged in any solution. It has been argued that this reduces the problem considerably to one that has a number of potential solutions [10].

1.2 The Clean Audio Project

The Clean Audio Project, funded by the Independent Television Commission (ITC) and Ofcom in the UK, investigated a number of approaches in collaboration with broadcast industry partners [11] in order to make use of the additional audio, and accompanying metadata about the audio, in 5.1 surround sound broadcast to improve television sound for people with hearing impairments. The most widely adopted of these potential solutions took advantage of mixing guidelines that recommend that speech is placed in the center speaker of a 5.1 reproduction system. The solution recommended was that, where appropriate, channels other than the center (normally speech) channel would be attenuated by simple switching, or by variable user-control. Where attenuation of non-center channels would clearly be inappropriate, for example for music broadcast and (comparatively rare) occasions where speech was placed in other than center channel for creative reasons, broadcast metadata should be used to flag the media as inappropriate for a clean audio mix. This use of broadcast metadata is highlighted in Fig. 1 as submitted to the Open IPTV Forum from the UK Clean Audio Forum [12]. In this reference implementation left surround, right surround, and LFE channels were excluded from the clean audio mix as no broadcast content had speech content in those channels.

Recommendations from this research, illustrated in Fig. 1, have since been standardized in ETSI TS101154 [13] and are referenced in EBU [14], Open IPTV Forum [15], and Nordig [16] standards.

Further experimental work carried out indicated that this solution was also effective at improving speech clarity when the resulting audio channels were downmixed for two-channel stereo reproduction [17] thus providing a solution requiring no additional bandwidth and with minimal additional processing requirements at the set top box. The solution is limited however by the inherent variability in the complex nature and degree of hearing loss experienced by individuals; there is unlikely to be a single, one-size-fits-all solution that works well for all people's hearing loss.

Since the recommendations provided by the Clean Audio Project were published there have been some further important research developments in broadcast audio. It is likely that, in much the same way that the growth of 5.1 surround sound audio facilitated the solution currently standardized, developments in object-based audio can provide further personalization of broadcast audio that will be useful in providing appropriate audio content for people with hearing impairments.

2 OBJECT-BASED AUDIO

Currently the most common use of object-based systems is in spatial audio implementations and the spatial audio scenario is used to describe the principles of object-based audio here. Using object-based audio systems for media personalization and to generate clean audio for hearing impaired viewers is then discussed.

Channel-based audio, as is currently standard across all mainstream broadcast platforms, is always defined with reference to a specific loudspeaker configuration; 2.0, 5.1, 7.1, etc. A sound is panned to a specific loudspeaker, or between two loudspeakers, and there is an underlying assumption that the same configuration will be present at production and reproduction parts of the broadcast chain. Clearly for broadcast this is not always the case and downmixing algorithms have been standardized in order to allow, for example, 5.1 media to be reproduced for two-channel stereo or mono reproduction. Nevertheless the audio channels that are broadcast and received at the set top box are, in reality, loudspeaker feeds based on the loudspeaker configuration on which the program material was mixed.

In an object-based audio system this convention no longer applies and each individual sound event within a sound scene can retain separation from the remainder of the sound scene. The audio object is defined by metadata that describes the sound event's characteristics, such as spatial coordinate position and other attributes. Thus object-based audio is considered to be *speaker layout format agnostic* [2] and can be rendered to any configuration of loudspeakers, as long as the object-based renderer is aware of the reproduction system being used. In practice, and in order to maintain compatibility with existing media and production workflows, object-based formats provide for channel-based *beds* that have audio objects added to them.

Object-based audio opens up a number of interesting possibilities for producing personalized audio content specific for hearing impaired people, or indeed for viewers' specific individualized hearing loss. Experiments carried out by Fuchs et al. [18–21] showed some of this potential using current broadcast technologies during broadcasts of tennis from the Wimbledon Tennis Championship. In the experiments carried out audio was mixed using object-based methods and unmixing metadata was transmitted over IP networks simultaneous with the resultant stereo mix. This allowed speech to be separated from other sounds at the decoder and for viewers to adjust the relative levels of on-court ambience and commentator according to their preference.

Research presented as part of the EU FP7 FascinatE project [22] also illustrated the potential for object-based clean audio for live events [23]. In the FascinatE project on-pitch sounds from an English Premier League football game, such as ball kicks and whistle blows, were captured as audio objects and transmitted to the reproduction system with positional metadata separate to crowd sounds and commentary streams. During the final demonstration of the project levels of on-pitch sounds, crowd noise and commentary could be independently controlled although formal assessment of this element of the project was not carried out at that time.

More recently Jot et al. [24] described an object-based system utilizing audio object loudness metadata to dynamically alter dialogue loudness based on personalization and non-dialogue program loudness. This development utilizes personalized clean audio by maintaining program level difference between *dialogue* objects and *non-dialogue* content. However for many genres it can be argued that some non-dialogue content also has an important role to play in making narrative comprehensible (e.g., the sound of a door opening as the protagonist enters the room, his or her footsteps as they approach out of camera shot). This suggests that a taxonomy of objects based on narrative importance and producer intent could generate more intelligent personalized audio than is possible using binary dialogue/non-dialogue definitions. Elicitation tests were carried out as part of the S3A project [25, 26] in order to better understand audio object perception and the results of these tests have been utilized in the research documented in this paper.

3 AUDIO OBJECT CATEGORY ELICITATION

Categorization is a fundamental process in human cognition, its main function being to reduce the volume of information processing needed to make sense of the environment. Object-based audio opens up the possibility of object level manipulation of audio content to optimize listener experience at the client end of the broadcast chain. Therefore, knowledge of the perceptual categorization of broadcast audio objects in complex auditory scenes will be of benefit to the development of object-based audio.

As part of a larger set of experiments on the categorization of broadcast audio objects, carried out as part of the S3A project, a free card sorting experiment was carried out with the aim of determining cognitive categories for audio

objects in feature film program material. The results of that experiment have helped to inform the object categories used for tests described later in this paper.

3.1 Method

In this experiment, 21 participants were presented with 8 clips from the film *The Woman in Black* [27]. The length of the clips ranged from 70 s to 223 s. Audio was reproduced using Genelec 8030A active loudspeakers arranged in a 5.0 setup in accordance with ITU-R BS.775-3 [28]. Video content was reproduced via a laptop with a 15.6″ screen (1366×768 resolution) positioned on a table in the test room approximately 0.8 m from the participant. The experiment was conducted in a semi-anechoic chamber at the University of Salford.

Participants were asked to sort a set of cards containing every identifiable sound in the clips into groups (142 cards in total), such that the cards in each group served a similar purpose in the composition of the scene. Participants were allowed to form as many groups as they wished. Once the participant was happy with their grouping, they were asked to give a label to each of the groups they had formed. In total, 110 groups were formed with the median number of groups formed by each participant being 5.

A matrix of category labels and audio objects was built that took on a value of 1 if an object had been grouped in a given category and a 0 otherwise. This matrix was subject to hierarchical agglomerative clustering (Ward method [29]) to investigate the structure of the participants' groupings. The results of this analysis were visualized using dendrograms, which show the categorization structure at different levels. In the full clustering solution, the bottom level of the dendrogram shows each object as an individual cluster; at each subsequent level of the dendrogram, the closest pair of clusters are merged until, at the top level, a single cluster is formed. Clusters can then be formed by cutting the dendrogram at a predefined level.

The resulting clustering solution was cut so as to result in five clusters. A five-cluster solution was chosen as this is the median number of groups that participants formed. The labels associated with each of the resulting clusters were interpreted by the researcher (see [26] for more details on this process). A dendrogram showing the resulting category labels is shown in Fig. 2.

The first of the five categories was interpreted as sounds related with actions and movement and related to foreground effect objects. The second category was interpreted as clear speech and related to dialogue objects. The third category was interpreted as prominent attention grabbing sounds and related to high-level transient foreground effects. The fourth category was interpreted as non-diegetic music and effects. The fifth category was interpreted as background sounds and included both discrete background effects and continuous diffuse background objects. These object categories were then adapted for the interface used for media personalization and simplified in order to present an easy to understand menu interface with intuitive categories. The non-diegetic music and effects category was

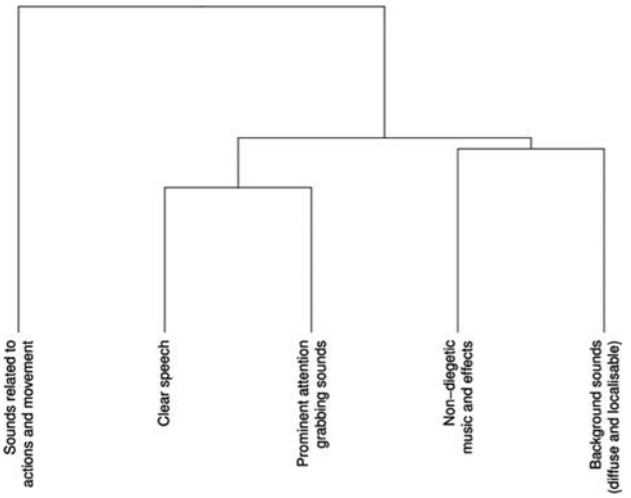

Fig. 2. Dendrogram showing hierarchical agglomerative clustering of category labels.

simplified as "music"; non-diegetic effects being unusual and usually considered as part of the production score. "Sounds related with actions and movement, and related to foreground effect objects" was labelled as "foreground effects" and tended to include sounds that were diegetic and carried some narrative meaning. "Background sounds" was labelled as "background effects" reflecting the types of sounds in this category, and "Speech" remained as a separate category. The category of "prominent attention grabbing sounds" was the least well defined of the categories arising from the feature film material when objects across different program material types were analyzed together. These were considered in the simplified category "foreground effects" in order to present a simple and intuitive user interface for testing.

Hence four audio object test categories were defined that would readily be recognized and understood by test participants: speech, music, foreground effects, and background effects.

4 MULTI-DIMENSIONAL AUDIO (MDA) CLEAN AUDIO IMPLEMENTATION

An implementation of clean audio using audio objects was carried out utilizing object-based audio tools in order to assess the preferences of hearing impaired people when presented with an interactive experience using audio objects. The production of the clean audio mix used professional industry tools based on the open standard known as MDA (Multi-Dimensional Audio) [30]. Using a content creation tool for cinematic immersive sound the soundtracks of the test films were post-produced to create stems that matched the object categories defined earlier. The stems were then turned in audio objects. These objects carried descriptive metadata, including object ID, object type, gain value, and three-dimensional position with (appropriate panning data). The final mix was output as an MDA bitstream, which then fed into the playout system. As part of the playback

Table 1. Details of each of the clips used in the test

Clip	Length	Video	Audio	Genre
Cold Calling	1:08 mins	HD 1080p	48 kHz, 32 bit.	Drama
Elephants Dream	1:43 mins	HD 1080p	48 kHz, 32 bit.	Animation
Exec-Corsist	0:55 mins	HD 720p	48 kHz, 32 bit.	Drama
Football (FASC)	1:33 mins	HD 1080p	48 kHz, 32 bit.	Sport
Never Forget	1:21 mins	HD 1080p	48 kHz, 32 bit.	Drama

experience, an intuitive on-screen GUI was presented to the participants to allow them to individually control the volume level of each of the audio objects in the mix. The user interface was designed with simplicity in mind, allowing for metadata driven customization (via on-screen menus), and an easy to use remote control.

5 TEST METHODOLOGY

5.1 Aims

The tests had two aims:

- Ascertain usefulness of personalization based on audio object categories for hearing impaired people including the usability of a screen-based interface;
- Understand the variation of level preference across people with differing hearing impairments.

5.2 Test Material

During the test process five clips with separate stereo sound stems for the four audio categories were used. The clips were between 55 seconds and 1:43 minutes in duration and of mixed genre including drama, animation, and sport with varied dynamic range and audio category content.

Clips sections were selected using a number of criteria:

- Combination of dialogue, music, background sound effects, and foreground sound effects (connected to on-screen events) occurring simultaneously;
- Length of between 1 minute and 2 minutes.

Audio was assigned to object categories by the sound designer of the films based on the descriptions described above and discussions about the perceptual attributes previously identified. The final choice of category allocation, which is at least in part a creative process, was taken by the sound designer based on his understanding of the producer's intent and of the object-categories.

Initial playback levels for each object category (before user adjustment) remained at the same levels as in the original 5.1 or stereo mix of the film thus presenting test participants with the same initial mix as was intended to be heard by the producer and sound designer.

After a short "training" phase during which participants could experiment and become familiar with the interface, each test clip was played repeatedly and participants encouraged to set preferred levels of each audio object category using on-screen menu level controls for each category. Once the participant had indicated that they had set levels of each object stream category to their preference the test moved on to the next test clip and the levels for each category logged in a text output file. The order of test clips was randomized between participants to avoid biases caused by increased familiarity of the interface.

Table 1 shows details of each clip, followed by a brief synopsis.

Cold Calling (Dan Price, Director):

Synopsis: A mysterious girl traps a Cold Caller.

Category contents:

- SPEECH: Conversation, whispering;
- FGFX: Diegetic sound; Gate noise, falling objects;
- BGFX: Non diegetic; Traffic noise;
- Music/ambience: Minimal.

Elephants Dream (Bassam Kurdali, Director):

Synopsis: Two animated characters move around an imaginary "machine," which seems intent on causing them harm.

Category contents:

- SPEECH: Conversation, shouting;
- FGFX: Diegetic; Lights buzzing, footsteps;
- BGFX: Tentacles, concrete grinding;
- Music/ambience: Strings, low to moderate volume.

Exec-Corsist (Dan Price, Director):

Synopsis: An Exec-Corsist is hired to "Exec-Corsize" the reality TV from a TV addict.

Category contents:

- SPEECH: Conversation, shouting, groaning;
- FGFX: Rattling lights, tables moving (on camera);
- BGFX: Tables rattling (off camera);
- Music/ambience: Pipe organ; Moderate volume.

Football (live capture)

Synopsis: Live footage from Match of the Day. Recorded as part of the FascinatE Project.

Category contents:

- SPEECH: Match of the Day commentary;
- FGFX: Audio objects: ball kicks, referee whistle that were isolated from crowd and other sounds;
- BGFX: Non-specific on field noise picked up by pitch-side microphones;
- Music/ambience: Crowd noise from crowd microphones.

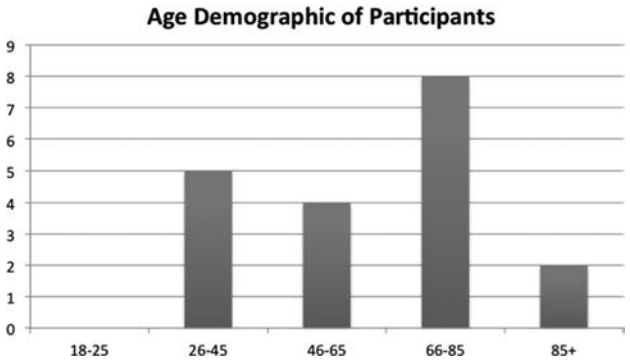

Fig. 3. Age demographic of participants in tests.

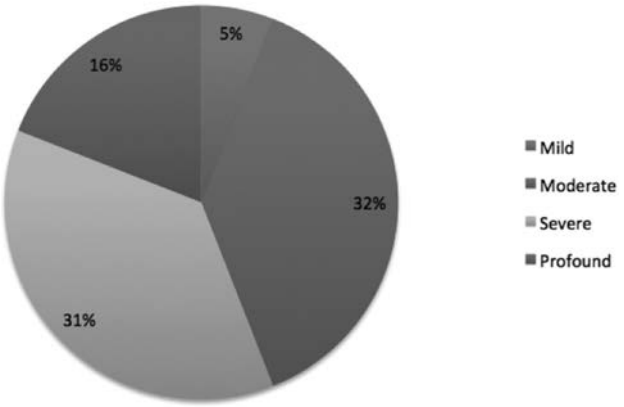

Fig. 4. Hearing loss of hearing impaired participants based on pure tone audiogram (PTA) [32].

Never Forget (Dan Price, Director):

Synopsis: A husband takes permanent revenge on his unfaithful wife.

Category contents:

- SPEECH: Conversation, phone call;
- FGFX: Footsteps, mobile phone ringing;
- BGFX: Traffic noise;
- Music/ambience: Club scene, loud.

5.3 Analysis of Participants

A total of 19 participants completed assessments, with 14 individuals in the group having some degree of hearing impairment based on *British Medical Journal* definitions of hearing loss [31]. Five participants with no reported hearing loss also took part in the tests.

5.4 Experimental Method

Testing took place in a listening room at University of Salford conforming to ITU-R BS.1116-3 [33], participants took part in tests individually. Test times took around 25 minutes per participant.

Audio-visual media was reproduced with video presented on a HD television and audio over a pair of full range loudspeakers positioned 30° either side of center at a height of 1.2 m as per ITU recommendation ITU-R BS.775-3 [28]. Participants were instructed on the test requirements

Fig. 5. Player interface enabling personalization based on four elicited object-stream categories.

and introduced to a simple intuitive user-interface that they would be using for the test. The object-based media player had a metadata driven, customizable on-screen menu system that allowed participants to alter the master volume, and the individual loudness levels of each object in the object-based mix that used the modified categories based on those elicited as part of the S3A project (described earlier and shown in Fig. 2). Once participants could confidently navigate the menu and alter volume levels for different object-category streams the test commenced. Each participant used a remote control to set the master volume for each media clip during the tests.

Clips were presented in random order for each participant to reduce learning and order effects.

Once the participant was satisfied with the levels they had set for each category for "the best clarity and understanding of the on screen action" a log file was generated of all the saved settings by the media player that were used later for data analysis. The log file recorded loudness levels for each category with reference to the default set by producer. Following the listening experiment, participants took part in a brief questionnaire to gather data on the ease of use of the test interface.

Owing to a technical issue one participant had incomplete results in the log output file so their listening test results are excluded. However results for the questionnaire for this participant are included.

6 EXPERIMENTAL RESULTS

Experimental results for all hearing impaired participants are presented here. Five non-hearing impaired people also took part in the tests, however the small number of these participants means that statistical data analysis for this group was unlikely to render any useful results so results are not presented here.

In the figures shown in this section, "BGFX" refers to background sound effects not connected with on-screen events, "SPEECH" refers to all speech content, "FGFX" refers to foreground effects, connected with visual events on-screen, and "MUSIC" refers to all music content.

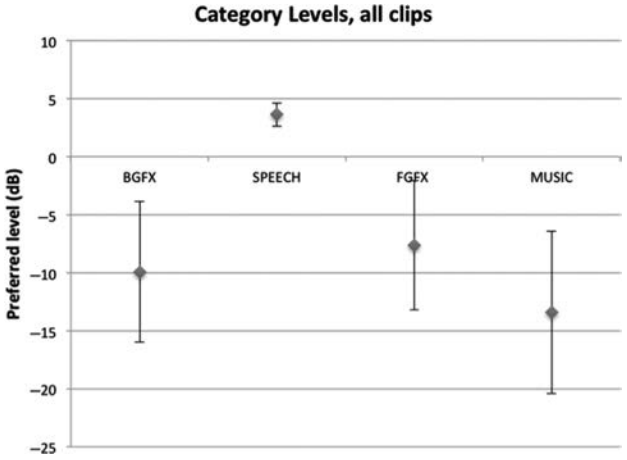

Fig. 6. Mean preferred levels for each sound category, 0dB represents the default level set by the production mixer, error bars indicate 95% confidence intervals.

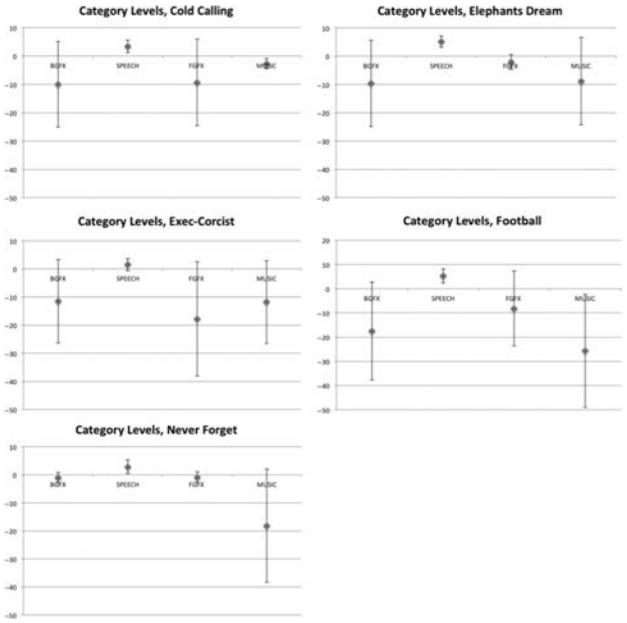

Fig. 7. Mean preferred levels for each sound category for individual media clips, 0dB represents the default level set by the production mixer, error bars indicate 95% confidence intervals.

6.1 Results for Hearing Impaired Participants

The overall mean levels for each category for all clips and error bars indicating 95% confidence levels are presented in Fig. 6. These data were analyzed using a repeated measures ANOVA (analysis of variance), which revealed a significant difference ($p < 0.05$) in the level of the speech/dialogue object compared to the other object types. No other statistically significant differences were found among the mean preferred levels for the other object types across all clips. It can be seen from this figure that there is much less variation in the mean preferred level of speech compared to the other object types.

Fig. 7 shows the overall mean levels and 95% confidence intervals for the individual clip types. The data from each individual clip type were analyzed using repeated measures ANOVA. For the Cold Calling clip, SPEECH was significantly higher than all other categories ($p < 0.05$). For the Elephants Dream clip, SPEECH was significantly higher than FGFX ($p < 0.05$). For the Exec-Corcist, Football, and Never Forget clips, no statistically significant differences were found between the mean preferred levels set for the different object categories.

6.2 Questionnaire Results

After taking part in the level-setting tests participants were asked questions via a questionnaire to discover if the interface was useful and/or appropriate to their needs. The frequency of responses across both hearing impaired and non-hearing impaired participants for each of the questions are shown in Table 2.

Additional comments added to the questionnaires were as follows:

Participant 3:

"I think this is a really good idea, simple but effective."

Participant 5:

"Useful and interesting to be able to adjust different aspects of sound."

Participant 8:

"It would be good to have a trial at home."

Participant 9:

"I would love to have access to this."

Participant 11:

"Missed subtitles."

Participant 13:

"First time I have been able to understand dialogue without subtitles in a very long time, I really liked being able to control the sounds."

Participant 14:

"Very straightforward, very good, when can I have one?"

Participant 15:

"Needs subtitles or wouldn't watch."

7 DISCUSSION

It is clear from the results shown here that, as assumed in previous studies [11, 18–20, 24], speech/dialog level was considered to be the most important feature and was

Table 2. Frequency of questionnaire responses for hearing impaired participants

Question or statement	Strongly disagree	Disagree	Neither Agree nor disagree	Agree	Strongly agree
1) I found the system easy to use.	1		1	5	8
2) I found it useful to be able to change the levels of sound separately.			1	5	9
3) I found the testing process easy.		1		7	7
4) I found the on screen menu easy to use.				8	7
5) I found it easy to use the remote control setup.				8	7
6) I found the foreground effects important to understand the story or on-screen action.	2	6	3	1	3
7) If I had this system at home I would leave the settings as I prefer them for all programs.	2	4	4	2	3
8) I watch a lot of television.		4	2	7	2

consistently set higher than other object categories by hearing impaired participants. Additionally 4 out of 15 hearing impaired participants stated in the questionnaire that they "found the foreground effects important to understand the story or on-screen action" indicating that, for at least some hearing impaired people, allowing greater than binary personalization (speech vs. non-speech) over program content was beneficial. Four participants also consistently set foreground effects higher than background effects in the tests. This preference for non-binary personalization, however, was only for *some* hearing impaired participants whereas speech preference was consistent across all.

A clear feature of the listening test results is that of a considerable variation in preference between participants for all factors other than speech as shown by the confidence intervals in Fig. 6. The low number of participants in each hearing impairment category (mild, moderate, severe, profound) makes detailed analysis of any link between hearing impairment category impossible, however it seems likely that factors other than pure tone hearing acuity, such as spectral and temporal resolution, also play a substantial role in narrative comprehension of AV media.

There was some, although much less, variation between media clips for each participant. This raises important questions for personalization for media content and, in particular, the personalization user interface. If the approach to personalization user interface design is done well, like with the media player used, and where personalization setup was a one-time event, or even per genre, i.e., different settings for drama, sport, etc., this could be considered as an acceptable overhead. However if personalization was required per program it would very likely be considered as unacceptable for TV viewers at home. Variation within subject choices may of course be simply a result of requiring each media clip to be personalized on an individual basis.

The personalization interface was considered by 13 out of 15 participants to be either *easy* or *very easy* to use, an important consideration when designing user interfaces for older people. Four categories of audio objects were clearly not considered to be excessively onerous to adjust by the participants.

In implementing any such system for broadcast unhelpful variation in personalization requirements may also occur if object categorization is inconsistent across programs. For 4 out of 5 of the media clips used here the sound designer was the same person and there was a clear understanding of the context of the grouping used from discussions within the research team. In order to take this out of the lab and into a broadcast environment a shared taxonomy for audio objects would have to be in place. A further question then is whether such a taxonomy, based on what is important to comprehension for a hearing impaired person, is the same as a taxonomy for narrative importance based on producer-intent.

Tools exist today to take the next steps. MDA allows for object-category metadata to be authored into the audio bitstream. If object-category hierarchical metadata is used at production, there is the potential to present various user interfaces. One consideration is by using an interface such as that used for these tests, the interface was well received and considered not overly complex for the task. Alternatively the object-categories presented here could be combined with the method documented by Jot et al. [24] in setting dynamic offsets based on multiple audio-object categories, instead of using a binary speech/non-speech definition.

There is also usefulness beyond personalization for hearing impaired people. The same method could be useful for visually impaired people as well, as useful non-speech audio object categories could be useful in helping to understand narrative of on-screen events. Hearing the protagonist's footsteps as he/she crosses the set and the door opening before a character walks into the scene could all help make drama narrative clearer. This will be explored in future tests. Further, following the useful principal that inclusive design is often better design, the same personalization as carried out here could be useful for people with no sensory

impairments experiencing media in sub-optimal situations, for example in environments with high background noise or on poor reproduction equipment. To this end future work is planned to include non-hearing impaired participants in varying acoustic environmental conditions.

8 CONCLUSIONS

Media personalization, to improve television sound for hearing impaired people, has been demonstrated using MDA, an open industry format for object-based audio. Object categories, elicited from perceptual clustering using free card sorting exercises, were used to create audio objects and presented as interactive menu items allowing individual level control of each category by hearing impaired participants. Results suggest that an individualized clean audio profile (instead of a one-size-fits-all) may be most effective for accessible television audio. The work presented also suggests that multi-category personalization can present a good solution for some hearing impaired people. Further work is identified to implement object-based personalized audio for visually impaired people and for sub-optimal listening environments.

9 ACKNOWLEDGMENTS

The object categorisation work documented here was supported by the EPSRC Program Grant S3A: Future Spatial Audio for an Immersive Listener Experience at Home (EP/L000539/1) and the BBC as part of the BBC Audio Research Partnership. Other experimental work was supported by DTS, Inc. using MDA and DTS:X technologies.

10 REFERENCES

[1] S. A. Silva, *Object-Based Audio for Televison Production* (2015).

[2] B. Shirley, et al., "Platform Independent Audio," in *Media Production, Delivery and Interaction for Platform Independent Systems: Format-Agnostic Media*, O. Schreer, et al., Eds. (Wiley: UK, 2014).

[3] ITU, *Sound Systems for the Hearing Impaired* (1994, ITU).

[4] C. D. Mathers, *A Study of Sound Balances for the Hard of Hearing* (BBC, 1991).

[5] A. Carmichael, et al., "The Vista Project: Broadening Access to Digital TV Electronic Programme Guides," *PsychNology J.*, vol. 1, no. 3, pp. 229–241 (2003).

[6] J. G. Beerends and F. E. De Caluwe, "The Influence of Video Quality on Perceived Audio Quality and Vice Versa," *J. Audio Eng. Soc.*, vol. 47, pp. 355–362 (1999 May).

[7] D. J. Meares, *R&D Report 1991–14: HDTV Sound: Programme Production Developments* (BBC: BBC London, 1991).

[8] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent Component Analysis and Applications* (Academic Press, 2010).

[9] M. Armstrong, *Audio Processing and Speech Intelligibility: A Literature Review* (BBC: London, 2011).

[10] B. G. Shirley, "Improving Television Sound for People with Hearing Impairments," in *Computing, Science & Engineering* (University of Salford: Salford, UK, 2013), p. 222.

[11] B. G. Shirley and P. Kendrick, "The Clean Audio Project: Digital TV as Assistive Technology," *J. Tech. & Disability*, vol. 18, no. 1, pp. 31–41 (2006).

[12] UK Clean Audio Forum, *Liaison Statement from UK Clean Audio Forum to ITU FG IPTV*, International Telcommunications Union Focus Group on IPTV: Mountain View, USA (2007).

[13] ETSI, "ETSI TS101154 v1.9.1 Digital Video Broadcasting (DVB); Specification for the Use of Video and Audio Coding in Broadcasting Applications Based on the MPEG-2 Transport Stream," in *Annexe E.4 Coding for Clean Audio SA Services,* (ETSI: France, 2009).

[14] EBU, "EBU – TECH 3333: EBU HDTV Receiver Requirements," in *7.3 Clean Audio* (EBU: Geneva, Switzerland, 2009).

[15] Forum Open IPTV, "OIPF Release 2 Specification Volume 2 – Media Formats," in *8.2.3 Clean Audio* (Open IPTV Forum: France, 2011).

[16] NorDig, "NorDig Unified Requirements for Integrated Receiver Decoders for Use in Cable, Satellite, Terrestrial and IP-Based Networks v2.4," in *6.2.4 Clean Audio* (NorDig, 2013).

[17] B. Shirley, "Improving Television Sound for People with Hearing Impairments," in *School of Computing, Science & Engineering* (University of Salford: Salford, UK, 2013), p. 222.

[18] H. Fuchs, S. Tuff, and C. Bustad, "Dialogue Enhancement—Technology and Experiments," in *EBU Technical Review*, M. Meyer and L. Vermaele, Eds. (EBU: Geneva, Switzerland, 2012).

[19] H. Fuchs and D. Oetting, "Advanced Clean Audio Solution: Dialogue Enhancement" (IBC: Amsterdam, Netherlands 2013).

[20] H. Fuchs and D. Oetting, "Advanced Clean Audio Solution: Dialogue Enhancement," *Motion Imaging J. Motion Imaging J.*, (2014).

[21] J. Paulus, et al., "MPEG-D Spatial Audio Object Coding for Dialogue Enhancement (SAOC-DE)," presented at the *138th Convention of the Audio Engineering Society* (2015 May), convention paper 9220.

[22] Research Joanneum, et al., *FascinatE* (European Commission: Europe, 2010).

[23] B. Shirley and R. Oldfield, "Clean Audio for TV Broadcast: An Object-Based Approach for Hearing-Impaired Viewers," *J. Audio Eng. Soc.*, vol. 63, pp. 245–256 (2015 Apr.)

[24] J.-M. Jot, B. Smith, and J. Thompson, "Dialog Control and Enhancement in Object-Based Audio Systems," presented at the *139th Convention of the Audio Engineering Society* (2015 Oct), convention paper 9356.

[25] A. Hilton, et al., *S3A: Future Spatial Audio for an Immersive Listener Experience at Home* (EPSRC: UK, 2013).

[26] J. Woodcock, et al., "Categorization of Broadcast Audio Objects in Complex Auditory Scenes," *J. Audio Eng. Soc.*, vol. 64: pp. 380–394 (2016 Jun.)

[27] J. Watkins, *The Woman in Black,* Momentum Pictures (2012).

[28] ITU, *ITU-R BS.775-3, Multichannel Stereophonic Sound System With and Without Accompanying Picture* (ITU: Geneva 2012).

[29] J. H. Ward Jr., "Hierarchical Grouping to Optimize an Objective Function," *J. Amer. Stat. Assn.*, vol. 58, no. 301, pp. 236–244 (1963).

[30] EBU, *MDA; Object-Based Audio Immersive Sound Metadata and Bitstream* (ETSI France, 2015).

[31] British Society of Audiology, "Descriptors for Pure Tone Audiograms," *British J. Audiology*, vol. 22: p. 123 (1988).

[32] B. S. Audiometry, *Pure-Tone Air-Conduction and Bone Conduction Threshold Audiometry With and Without Masking* (British Society of Audiology).

[33] ITU, *ITU-R BS.1116-1: Methods for the Subjective Assessment of Small Impairments in Audio Systems including Multichannel Sound Systems* (International Telecommunication Union, 1997).

APPENDIX A: EXPERIMENTER SCRIPT

<Subject enters>

– Hello and thank you for attending today.

<Subject sits>

– The experiment is concerned with investigating the potential impact, if any, of using new audio technologies to enhance the listening experience of hearing impaired people whilst watching television or other broadcast media.
– When the test starts, you'll be shown 5 video clips on the television set. The sound on these videos will have been separated into four categories: Dialogue, Music; Foreground sound effects (sound which is related to something happening on screen – such as a telephone ringing or a door slamming) and background sound effects, such as traffic noise or people chatting in the background of a pub scene.
– At the start of the video the four categories will all be at equal volumes, and can be adjusted separately using the remote control provided.
– Please "mix" the volumes of the four categories to the levels which give you the best clarity and understanding of the on screen action, as if you were at home watching your TV or listening to the radio.

<Hand subject the remote>

– You can adjust the levels of the four audio categories by using this remote control, which allows you to adjust the level of each audio category separately. For example, if you wished to adjust the volume of the music or dialogue, you would turn the channel marked "MUSIC" or "?DIALOGUE' up or down.

<Demonstrate>

– The videos will loop to give you plenty of time to adjust the levels to the way you want them. If by the end of the video you are happy with the levels, please move on to the next video. Alternatively, wait, and the video will play again.
– If you have any problems, please ask for assistance.
– Do you have any questions?
– Are you ready to begin?

THE AUTHORS


Dr. Ben Shirley


Melissa Meadows


Fadi Malak


Dr. James Woodcock


Ash Tidball

Dr. Ben Shirley is a senior lecturer in audio technology at the Acoustics Research Centre, University of Salford, UK. He received his M.Sc. from Keele University in 2000 and his Ph.D. from the University of Salford in 2013. His doctoral thesis investigated methods for improving TV sound for people with hearing impairments. His research interests include audio broadcast, spatial audio, and also audio related accessibility solutions for people with sensory impairments.

•

Melissa Meadows has an M.Sc. in digital media from the University of Salford and a Master of Arts Degree in television production from the University of Manchester. Melissa is interested in acoustics and psychoacoustics research, specifically in the potential of multichannel system effectiveness for immersion and envelopment techniques in theater. Research has included the human auditory system, broadcast technologies, and an international collaboration between the University of Salford and DTS, researching and testing the development of an object-based clean audio system for hearing impaired people.

•

Fadi Malak is responsible for defining the strategy and go-to-market plan for DTS's next-generation platforms. His current focus is on the broadcast market strategy and DTS's new object-based audio solutions, which includes heading strategic partnerships and international standards work. Fadi began his audio career at Dolby, where he spent over four years helping Dolby break into new markets with both their audio and video technologies. A broadcast industry veteran, he also spent six years with Harmonic, helping them launch the H.264 compression market and IPTV. In the 1990s, Fadi began his A/V technologies career at Apple computer, as a multimedia technology evangelist, has written several papers, contributed to patents, spoken at numerous industry events, and participated as a member of the ITU SG6, DVB CM & TM, ATSC, SCTE, EBU, and SMPTE. He currently serves on the Board of the UltraHD Forum and VRIF.

•

Dr. James Woodcock is a research fellow at the University of Salford. His primary area of research is the perception and cognition of complex sound and vibration. James holds a B.Sc. in audio technology, a M.Sc. by research in product sound quality, and a Ph.D. in the human response to whole body vibration, all from the University of Salford. James' work is currently focused on how object-based audio can improve the listener experience of spatial audio.

•

Ash Tidball has worked in the film and TV industry for the past 15 years, primarily in sound and directing. On the one hand, he is a composer and sound designer of commercials, stage shows, and award-winning short films including BBC music documentaries and post-production sound for TV. He is also director and editor of independent films, several government mental health interest films, and has won several awards on both sides of the Atlantic. He teaches media and TV broadcast at the University of Salford.