Clean Audio for TV broadcast: An Object-Based Approach for Hearing-Impaired Viewers

BEN SHIRLEY, *AES Member,* **AND ROB OLDFIELD,** *AES Member*

(b.g.shirley@salford.ac.uk)   (r.g.oldfield@salford.ac.uk)

University of Salford, Salford, UK

A significant proportion of any population suffer from some form of hearing loss, and age demographics mean that this proportion is increasing. A prominent effect of hearing loss is an increased difficulty to pick out speech signals in the presence of background noise or music, which can be particularly problematic for broadcast audio content. Consequently a common subject of complaints to broadcasters is that the dialogue is masked by background noise/music or that speech intelligibility is poor. This paper presents these issues, reviews previously proposed solutions, and presents an object-based approach that can be implemented within MPEG-H that can give users control of their audio mix and hence facilitate enhancements to be made for hearing impaired viewers.

0 INTRODUCTION

There are estimated to be around 10 million people who are deaf or hard of hearing in the UK including around 8.3 million who suffer from mild to moderate deafness [1] and would gain some benefit from a clean audio provision in television audio. The term clean audio is used here to mean an audio mix with reduced background noise and/or enhanced dialogue. Support groups for people with hearing impairments—such as Hearing Link (previously known as Hearing Concern and Link) and national broadcast watchdog organizations such as the Independent Television Commission (ITC) and then Ofcom in the UK—receive many complaints from hard of hearing people that the dialogue on TV is unclear and hard to understand owing to the level of background "noise." This noise consists of background music, sound effects, and speech that can mask the dialogue thus reducing comprehension. This level of complaints to user groups and broadcast regulators has been mirrored in complaints to broadcasters. A survey carried out by the BBC in 2010 indicates that 60% of viewers had difficulty in understanding speech on TV [2]. Digital TV, and especially the increasing availability of surround sound content, has the potential for much improved TV sound quality and could therefore be of great benefit to hearing impaired viewers. However the wish to take advantage of the increased dynamic range available in digital TV broadcast may instead lead to increased problems for hearing impaired people. More recently there has been significant interest in a so-called object-based approach to audio with standardization activity taking place in the form of MPEG-H [3–5]. Object-based audio moves broadcasts away from the tradi-tional channel-based paradigm where the requisite number of audio channels are broadcast to match the reproduction system at the user end and, instead, broadcasts the necessary audio components needed in order to reconstruct the audio scene at the user end. Applying such an object-based approach affords much greater control over the reproduced sound scene and the balance between the various components such as speech, background noise, music, etc., and therefore shows much promise as a means of improving the audio for the hearing impaired population. This paper begins with an overview of the scale and nature of common hearing losses and then proceeds to discuss these effects in terms of broadcast audio content and how this is affected by digital broadcasts. Subsequent to this, an object-based approach to the problem is proposed as exemplified by the FascinatE Project and the standardization activity of MPEG-H. Finally a case study is given as to how such an approach could be practically realized for a real broadcast.

1 PREVALENCE OF HEARING LOSS

In order to more clearly understand the potential benefits, and also the constraints of any solution for clean TV audio that may work for the maximum number of people, it is useful to understand a little about the scale and nature of common hearing losses. This section outlines, as far as is known, the proportion of the population that may be helped by a clean audio solution. The number of people suffering from some form of hearing loss can be difficult to assess accurately. Action on Hearing Loss (AHL—previously the Royal National Institute for Deaf People (RNID)) estimate

the number of people suffering from hearing loss in the UK to be around 10 million [1]. Davis [6] carried out a survey across several cities in the UK and concluded that 16% of UK adults have a bilateral hearing impairment and 25% have a bilateral or unilateral impairment. Of these only 10% self-reported that they had bilateral hearing difficulty in a quiet environment indicating the difficulties of reliance on self-reported statistics. The Medical Research Council's statistics show clearly the correlation between hearing loss and age.

Even allowing for inaccuracies from self-reporting, the use of different classification systems across different countries has increased the complexities involved in understanding the prevalence of hearing loss. Nevertheless, the data available makes clear that hearing loss as a result of the aging process is widespread. In addition to age-related hearing loss (ARHL), the national burden of hearing difficulties attributable to noise at work is considered to be substantial [8]. Several sources provide indications as to the prevalence of noise induced hearing loss (NIHL) in the UK, a Health and Safety Executive (HSE) sponsored report [9] collates data from a number of sources as follows. In 1995 14,200 people were in receipt of benefit for industrial deafness [10], however a Medical Research Council survey quoted by HSE [11] estimates the true number to be closer to 509,000; the discrepancy being mainly because the conditions needed in order to claim benefit include a high degree of hearing loss (>50 dB in both ears). Among young people the impact and cause of hearing loss, in particular NIHL, may be different, one substantial component being entertainment. Studies into hearing impairments from recreational activities [12, 13] have found noise levels substantial enough to cause some permanent hearing damage with repeated exposure across a wide range of activities. Detailed studies into musicians [14, 15], employees in the entertainment industries [16], and young people listening to music on headphones [17] largely indicate substantial impact of entertainment sound exposure on the hearing of younger people. The variety of hearing impairments present in the population as a whole therefore varies considerably according to the cause of impairments. Even within the population that suffers from ARHL each person's hearing loss is unique in terms of its frequency characteristics. Furthermore issues such as loudness recruitment (an increased perception of loudness) and frequency selectivity or frequency resolution (leading to spectral blurring and masking effects) mean that the problem of improving TV sound for the hearing impaired group is therefore not as straightforward as might be supposed.

2 HEARING LOSS AND TV BROADCAST

A limited amount of research has been carried out specifically focused on TV audio for people with hearing impairments, principally by the broadcast industry. This section documents and comments upon that research. The methodology and detailed results of the BBC's own large scale survey into hearing impairments and TV audibility are unpublished, however some key outputs have appeared in news releases [2]. One of the headline findings was that 60% of viewers "had some trouble in hearing what was said in TV programs." Background noise and background music accounted for 25% of these; other major factors being foreign language and dialects, poor diction, and speech being "too fast." The survey led to a series of guidelines and training materials for the BBC in order to alleviate problems as far as possible through improved production techniques [18]. Over a number of years RNID research published in their annual survey report has held background noise accountable for a higher proportion of problems with dialogue on TV than the BBC research suggested. Reports indicate that the number of people finding that background noise affected their ability to hear speech on TV rose from 83% of respondents in 2005 [19] to 87% in 2008 [19]. The problem was worse for older people with 88% of the over 65 age group reporting problems compared to 55% of those aged 16–24. Interestingly 45% of those surveyed who had no reported hearing loss also noted background noise as affecting their ability to hear speech [20] indicating that different mixes, rather than the use of subtitles, is more likely to be a useful solution for many. The ITU have questioned how TV sound could be made better for hearing impaired people [21] and studies have been carried out by the BBC in the UK. A BBC study by Mathers [22] responded to complaints from viewers over a number of years complaining about "background sound (e.g., audience laughter, crowd noise, background music, etc.)." Mathers' research carried out subjective testing where participants were presented with three different levels of background sound; a reference mix between speech and background thought appropriate by a BBC mixing engineer and mixes with background sound at –6 dB and +6 dB relative to this recording. Three-hundred-thirty-six participants in total were asked to assess the perceived intelligibility of speech in excerpts across a range of program material; however, only 25% of these were tested under controlled conditions. Results from the research were inconclusive and led to an expressed view that intelligibility was not highly dependent on background level and that very high or very low levels of background sound would be needed for a significant impact to be shown. A second BBC study by Meares [23] suggests that multichannel sound systems associated with HDTV could be utilized in providing a Hard of Hearing (HoH) channel, at least for programming where a discrete commentary channel exists, and potentially for other programming, but only where access to the original mixes were available. Meares suggests that hard of hearing listeners would benefit enormously from such a provision but identifies potentially problematic additional costs in broadcasting these HoH services. An additional HoH channel is identified as being an ideal solution by others [24], however there is no clarity as to how a clean dialogue channel could be derived from material that is already mixed, as is usually the case. Methods developed during work carried out by Uhle et al. [25] as part of the European project "Enhanced Digital Cinema" (EDCine, IST-038454) aimed to reduce background sounds in the center channel and so improve speech quality; however the methods used, based on single channel separation, had limited

success. Research on audio processing to improve intelligibility for TV sound can be found in Armstrong's useful literature review [26] that points out the considerable difficulties inherent in separating speech from competing noise. He concludes that, "audio processing cannot be used to create a viable 'clean audio' version for a television audience." Indeed the problems of blind source separation for audio are well documented for natural acoustic environments and are of limited effectiveness except in anechoic conditions [27]. However it seems likely that produced media such as a surround sound, or even stereo mix, where mixing parameters are not simply the result of source and microphone placement but are produced subject to known guidelines and conventions, may present a special case. Examples of this are documented later in this paper. Some research has been carried out aimed specifically at improving the intelligibility of speech on TV audio for hearing impaired people. Early work by Carmichael [28] on the DICTION project indicated that, at the time of the research, although signal processing could make speech sound clearer, it could not improve measures of objective intelligibility in terms of word recognition. Müsch [29] has argued that processing can still reduce the cognitive effort required for comprehension and has discussed algorithms aimed at decreasing listener effort and, as a consequence, improving intelligibility. Müsch explains that older listeners tend to prefer higher listening levels to younger listeners because of elevated hearing thresholds but also that the effect of loudness recruitment reduces the level at which listening becomes uncomfortable or painful. Consequently there is evidence of what Rumsey refers to as a reduced "window of comfortable dynamic range" [30] for older listeners. Müsch argues that the cognitive load caused by the mental processing used to filter out background sound and "clean up" the speech means that there is reduced attention for the higher level cognitive processing used to contextualize sentences and therefore fill any "gaps" caused by words not heard. He suggests that the problem for older people in understanding speech on TV is not usually cognitive impairment but is primarily of a sensory nature. Reduced frequency resolution affecting the recognition of speech formants is cited as one reason, another being the reduced ability of hair cells in the inner ear to detect phase effectively. The arguments here are that audibility is key and that signal processing may not be able to improve individual word recognition but may be able to reduce listening effort and therefore the cognitive load that may play a part in preventing comprehension for hearing impaired people. Others, cited by Carmichael [28], have argued that there are more factors at work than simply the sensory impairments themselves. Cervellera [31] points to age related degradation of nerve pathways as adding significant "noise" to perceived sounds; Stine et al. [32] and Rabbit [33] point to evidence that slower and less efficient cognitive systems resulting from the aging process also add to the problem. Certainly the combination of high dynamic range audio, competing background noise, reduced comfortable dynamic range, lack of frequency resolution, and other effects brought on by physiological change such as degraded nerve pathways and reduced/slowed cognitive

Table 1. Age distribution of self-assessed hearing loss as published in *International Journal of Audiology* [7].

Adults Aged	With mild, moderate, severe, or profound hearing loss
16–60	6%
61–80	47%
81 and over	93%

performance, explain why older viewers may find it difficult to understand speech on TV and the subsequent number of complaints received by Ofcom and television broadcasters.

2.1 The Impact of Multichannel Audio Broadcast

As discussed already, instances of produced media clearly provide a different set of issues compared to the usual challenges of blind source separation of audio. Indeed multichannel surround sound formats may in themselves bring intelligibility advantages for hearing impaired (and other) television viewers. One solution, adopted into some broadcast standards [34, 35] identifies a simple approach that takes advantage of the fact that most 5.1 program content uses the center channel primarily for speech (where speech is present). The Clean Audio Project [36] demonstrated significant improvements in speech clarity for people with hearing impairments by a simple re-balancing of channels in a 5.1 mix where dialogue was in center channel only; this principle is detailed in Fig. 1. The re-balancing was to be controlled by broadcast metadata indicating whether the material was appropriate for this "clean audio mix."

Some research suggests that there may be some benefits for television sound by the addition of a central loudspeaker, as is used in 5.1 surround sound systems, compared to a central "phantom" stereo image. Often, where both 5.1 surround sound and two-channel stereo broadcasts take place, only one 5.1 mix is carried out and an automated down-mix used for stereo broadcast. Increasingly though, the 5.1 mix is the only available broadcast format and down-mixing occurs at the set top box in the users' home. It is suggested by Dressler [38] that the down-mix process, whereby a 5.1 surround sound audio stream is converted for two channel playback, may distort the mix in such a way as to reduce intelligibility by altering "the subjective balance of the mix." Holman [39] suggested that the addition of a central loudspeaker made the material easier to understand although stated that this may not actually produce greater intelligibility. This effect, leading to an apparent difficulty in understanding, is a result of acoustical crosstalk [40] that occurs when two identical signals arrive at the ear with one slightly delayed compared to the other. The resultant comb filtering effect cancels out some frequencies in the audio. Other work has shown small but significant measurable improvement in intelligibility by utilizing a central loudspeaker for speech instead of a phantom center [41]. The comb filtering effect caused by crosstalk has been found to be problematic to the listening experience more generally, in particular for transaural reproduction,

Optional user control of
the mix of speech and
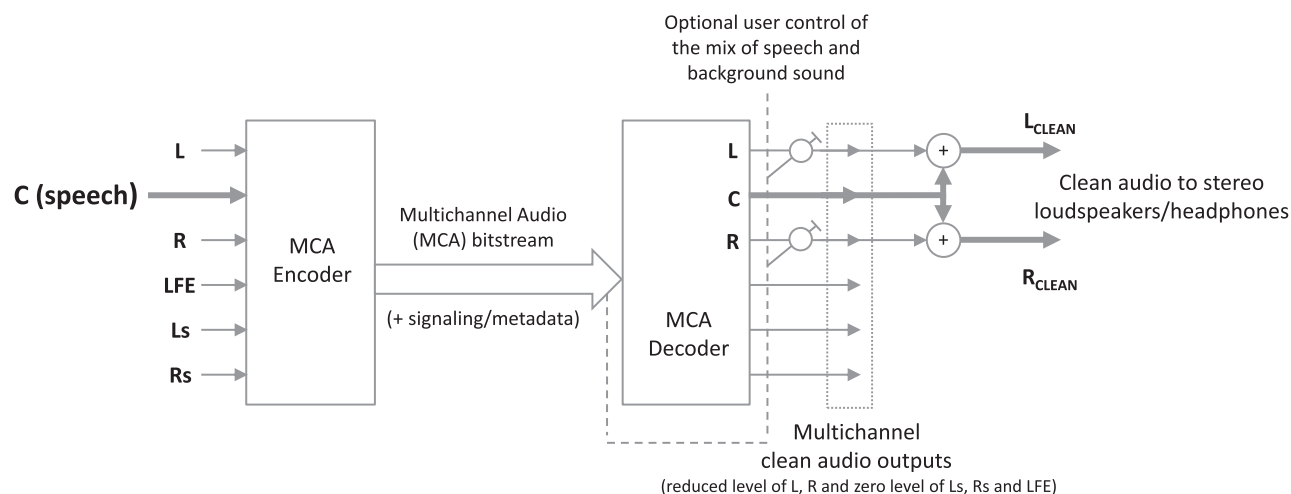background sound

Fig. 1. Example of a clean audio system reference model reproduced from UK Clean Audio Forum liaison statement to ITU Focus Group on IPTV [37]

and has therefore been the subject of considerable research. Commenting on frequency response problems caused by signal path delays David Clark [42] states that "clearly the 'phantom' center is an entirely different listening experience than pure left or pure right. One might ask if stereo is deeply flawed as [a] sound reproduction technique or if interference notches should simply be ignored." Impacts for listeners beyond intelligibility mean that considerable efforts have gone into attempts to remove, or reduce the effects of crosstalk, some of which could also have potential benefit for hearing impaired people in reducing the impact of crosstalk on intelligibility of speech noted in [41]. Methods have been proposed to reduce crosstalk impact by Cooper and Bauck [43] and [44] but these may be impractical in the context of broadcast television viewing as they utilize crosstalk cancellation techniques that rely heavily on the listener being in the ideal listening position. Clearly in a home environment this is very rarely the case. Vickers [45] recognizes this and points out that "when the listener is not equidistant from the speakers, the comb filter peaks and nulls will be in different frequency locations at the two ears." The resultant comb filtered perception in any given location in the room then becomes very unpredictable and impossible to compensate with additional comb filters. There is some debate as to the specific cause of intelligibility problems resulting from crosstalk. Bucklein [46] suggests that intelligibility difficulties may actually be made worse by peaks resulting from the crosstalk effect, rather than the troughs, as might be assumed; however the underlying problem remains, regardless of which effect of crosstalk is most detrimental. Other approaches to reduce crosstalk impact such as decorrelation methods [47, 48] have been suggested with the aim of randomizing the effects and to hence make the impact less prominent. These approaches can be seen as a signal processing equivalent of relying on room reflections to even a room response by creating a diffuse field, but some have cited artifacts and distortions from these methods that, with musical content, can manifest as unacceptable change in timbre [49]. Vick-

ers suggests a further possibility for improving the timbral quality of the center image (and therefore improving the intelligibility of centrally positioned speech); by deriving a center channel from two-channel stereo content, which would then be presented as a real, rather than a phantom, source. He suggests a method for accomplishing this using frequency domain upmixing [50] and provides a useful review of upmixing methods to this end [33]. Clearly this would be a useful direction for improved TV audio reproduction for hearing impaired listeners if it could be shown to be effective. Derived side channels (L and R) that did not contain speech could be reduced in level with reference to the new derived center channel content in order to improve speech intelligibility. However his research suggested that existing upmixing algorithms either provided inadequate center channel separation or produced "watery sound" or "musical noise" artifacts although formal subjective testing was not applied to assess this thoroughly [51]. In other related work Goodwin and Jot [52] make reference to primary-ambient decomposition in extracting ambient information from stereo signals using principal component analysis (PCA). Other research by Zielinski et al. [53] documents principal component analysis (PCA) processing to separate sources for produced media under certain specific conditions. The research utilized PCA using parallel signal paths as shown in Fig. 2.

The input for these experiments was in two parts; a reference 5.1 audio section taken from the media to be processed, carefully selected as containing speech, had a band pass filter applied to it discarding those frequencies that contained no speech components. PCA was carried out on the filtered reference audio and the principal component identified from the mix with this positive bias toward "speech frequencies." The second complete unfiltered input had PCA carried out on it and all components except the principal component (based on eigenvectors and eigenvalues that had already been determined by the filtered reference) were attenuated or deleted. The PCA process is therefore biased, or weighted, for speech frequencies as shown in Fig. 2.
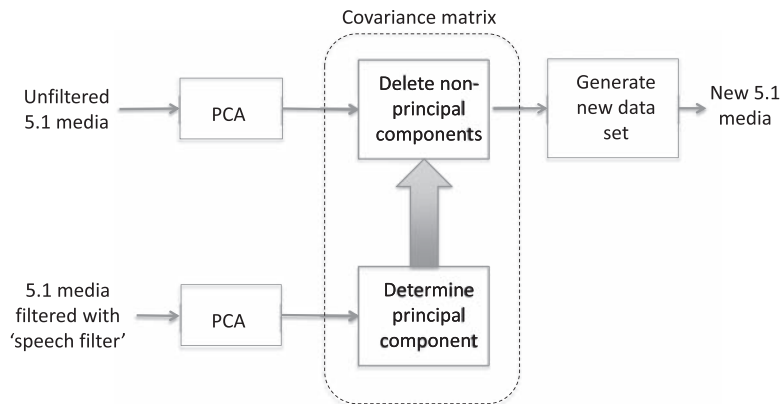
Covariance matrix

Fig. 2. System diagram showing a "speech filtered" covariance matrix determining attenuation or deletion of components in unfiltered path [54]

The method was shown to work well for situations where speech was in one channel only but had the disadvantage of requiring some considerable human interaction to take place before the process could be effectively carried out; the choice of section used to generate the unmixing matrix was key to the success of the method. Additionally the method assumes that the unmixing matrix generated by PCA of this section is applied to full duration of the 5.1 material, the assumption being that this will be appropriate for the entirety of the remaining media. The work presented in this review of prior research suggests that Armstrong's unequivocal conclusion that, "whilst audio processing can be used to create cosmetic improvements in a speech signal it cannot be used to improve the ability of an audience to follow the words" [26] may be valid for the general case, however actual intelligibility improvements have been reliably demonstrated under experimental conditions from produced multichannel media.

3 OBJECT-BASED AUDIO SOLUTIONS TO HI TV AUDIO

Broadcast and audio industry developments from Dolby and DTS [55, 56] have been at the forefront of recent commercial interest in object-based audio systems. Although initially aimed at redefining spatial audio production and reproduction, object-based broadcast has the potential to enable individually customizable audio experiences for viewers including for viewers with hearing impairments. In channel-based systems, such as those currently broadcast, audio event positions are defined in terms of the level and time relationships between the signals fed to the loudspeakers and are mixed into one "package" of channels for a specific target loudspeaker layout. Sound sources are mixed such that unmixing individual elements becomes a complex and often insoluble problem. Conversely, in an object-based system, audio events can remain separate and are tagged with coordinate (and other) information as metadata. While potentially making the produced audio mix less predictable on a "standard" reproduction system at the home, and hence less controllable by the mixing engineer, it does allow for a greater degree of flexibility in rendering at the set top box.

For program audio derived from multiple microphones and other produced media this has a number of benefits, one of which is that it can facilitate a format agnostic system [57] that is defined without reference to any specific loudspeaker configuration. As long as the decoder at the user's TV or set top box (STB) has an appropriate decoder, an object-based audio scene may be decoded to any loudspeaker layout without need for downmixing or upmixing. Recent developments by Fraunhofer have utilized object-based methods in order to create a "dialogue enhancement" algorithm that pre-prepares material for broadcast in such a way that it can be unmixed to two individual components at the set top box. In this implementation dialogue enhancement parameters are generated as part of the mixing process and these contain sufficient information to derive original speech sources from the mixed AV media at the STB. Instead of transmitting separate speech and background, an AAC or HE-AAC bitstream is transmitted that contains mono, stereo or 5.1 mixed content. Metadata containing the unmixing parameters required to separate out sources required to create a clean audio output are transmitted as part of this transmission. The advantage to this solution is that it could be made backwards compatible with existing equipment; where no decoder was present in the STB, the default stereo or 5.1 mix would be heard. This solution was demonstrated for two-channel stereo material as part of BBC Wimbledon coverage [58] where viewers were able to use a PC-based software application to adjust levels of commentary compared to court-side ambience. Although the process brought some additional complexity to the production process audience response was favorable. The technology is further described in [59] as an implementation of MPEG SAOC (Spatial Audio Object Coding) where the dialogue or commentary is considered as an audio object that can be extracted from the pre-mixed audio based on the parameters transmitted with the broadcast audio mix.

3.1 Object-Based Clean Audio in the FascinatE Project

The FascinatE project [60], completed in July 2013, was an EU FP7 project that developed a complete end-to-end future broadcast system designed to be format agnostic and

interactive [57]. The system was based around user navigation of an ultra-high definition panorama with accompanying adaptive 3D audio. The FascinatE project outcomes give an interesting view of how clean audio could be implemented in a future object-based audio system alongside other interactive elements. The premise of the clean audio work in FascinatE was that by retaining separation of speech, and other audio content important to the narrative meaning of the program material throughout the broadcast chain, it would be possible to implement a clean audio mix as suggested by the UK Clean Audio Forum [37] and implemented in EBU [34] and ETSI [50] recommendations. The standards suggest attenuation of "non-speech channels" by 6 dB (where speech is present and in center channel only); however by using an object based audio system it would be possible to set a personalized level of attenuation of speech and various non-speech objects depending on the requirements of the viewer. This section describes the project's implementation of clean audio and describes the techniques developed that would need to be adopted by the broadcast industry in order for these methods to become a reality. Here object-based audio was utilized as a means to provide dynamically matched audio for interactive navigation through an audio-visual scene. The project captured a very high-resolution panorama [61] (approx 7K × 2K pixels) and allowed pan, tilt, and zoom navigation of the panorama by the user. In order to provide matching audio for the user-defined scene it was necessary to move away from a conventional channel-based audio paradigm. Instead an audio capture system based on an object-based paradigm was used to capture the audio scene as part of a live broadcast; audio events were captured complete with 3D coordinate information specifying where in the audio scene the sound event had taken place. This is analogous to a gaming audio scenario and, in the same way as a first person game allows navigation around and between audio objects, the object-based audio capture enabled users to pan around and zoom into the AV scene with audio events remaining in their correct locations in the reproduced sound field. It was possible in the FascinatE system to zoom across a scene and past audio objects, which would then move behind the user's viewpoint thus realizing a realistic audio scene to match the chosen visual viewpoint. By utilizing the production and acquisition techniques developed in the FascinatE project a system was developed that could provide a clean audio output without any need to address the issue of separating speech components from the background "noise" that has been the subject of much of this paper. By maintaining a separation of all audio components and objects that make up a program throughout the broadcast chain it was possible to enable mixing of every aspect of broadcast audio at the user-end based on user preferences including a requirement for clean audio for hard of hearing people. For a produced program, such as that described in Forrester and Churnside's work on perceptive radio [62], it is relatively trivial to enable a hearing impaired mix to be made via simple user command; sound effects and music could be lowered relative to dialogue for example, and an implementation of 6 dB or any other attenuation of non-speech content

(even one customized based on user requirements) would be straightforward and could be readily adapted for accessibility purposes. The FascinatE project, however, focused on broadcast of live events and this created considerable challenges for the provision of object-based clean audio. The project covered a range of event genres as part of the test shoots and the example given here—that of live sports broadcast (see Sec. 3.3), is of most relevance to the concept of clean, or accessible, audio.

3.2 MPEG-H for Object-Based Clean Audio

The forthcoming MPEG-H standard [5] offers a practical implementation of an object-based audio system. In this standard audio elements can be defined and broadcast as separate components and the relative levels of each of these elements can be altered or panned within the reproduced sound scene at the user-end. MPEG-H supports channel-based systems as well, and audio elements can be either audio objects, ambisonic sound fields or loudspeaker channels. Using the MPEG-H format users can have controls over the individual elements of the broadcast sound scene including the selective muting or altering of individual channels. Due to advertising revenue streams in many countries it may not be possible to allow the reduction in level or muting of commentary channels but the MPEG-H standard also allows content creators to place restrictions on the interactivity/adjustment of the audio elements within a scene so this could be accounted for at the production end. In the context of this paper one could imagine the restriction being that the level of the commentary with respect to the background could only be raised rather than muted or lowered. Using MPEG-H there is also the possibility to include in the broadcast, commentary/audio from different languages making the broadcast more accessible.

3.3 Clean Audio for Live Sports Coverage

Consideration of providing clean audio for live sports coverage introduces an interesting question as to what audio should be considered useful and retained and what should be considered as background noise that may be detrimental for comprehension of the program material. The FascinatE project used the example of live football coverage as one of its test scenarios and this provides a good illustration of techniques developed, but, equally, these techniques could also be applied to other live event genres. Clearly, as in other program genres, speech (in this case commentary) is an important component to understanding the narrative and meaning of events during a football game. One could gain a clear understanding of what is happening on the football pitch by listening to the commentary channel alone, however other sound events, for example the sound of a referee's whistle, also provide meaning and narrative information. Sound design techniques, such as hyper-real and exaggerated ball-kick sounds that have become commonplace over the last 20 years, indicate that all of these on-pitch sounds are considered important to the experience of enjoying a televised football game. Indeed the exaggerated on-pitch

sounds introduced to live sport by Sky have been adopted in computer gaming and have become synonymous with live sports coverage. There is a parallel here with diegetic and non-diegetic sounds in film. Diegetic sounds are usually defined as "sounds whose source is visible on the screen or whose source is implied to be present by the action of the film" [63]. In improving the TV experience for hard of hearing people it may be that diegetic sounds that are critical to the narrative of the program should be treated differently to the background "noise" whose reduction has been the focus of much research cited here, thus it is important in an object-based system to capture and broadcast these sounds as separate elements to allow the users to customize their output to suit their individual needs. To this end three categories of sounds were considered: speech content whose comprehension is critical, background noise that has been shown to be detrimental to both clarity and to perceived overall sound quality, and other non-speech sounds that are considered important to comprehension and/or enjoyment of the material. In approaching an object-based audio broadcast these should each be capable of being processed or mixed independently either directly by the user, or based on predetermined user preferences at the set top box. In the example of a live football broadcast these categories consisted of speech in the form of live commentary, crowd noise that could be considered as detrimental to comprehension, and on-pitch sounds such as ball-kicks and the sound of the referee's whistle that may be important for comprehension and also for perceived overall sound quality. In current TV broadcasts these discrete audio object categories are not available at any point in the broadcast production chain and therefore the end user has no control over the relative levels of these sounds. In order to provide these three sound sources as independent and controllable entities some considerable development had to take place in the acquisition and production techniques used to capture a complex sound scene such as that found at a live sports event [64]. Currently the key objectives for audio in football coverage are twofold; picking up sounds on the pitch as clearly as possible during the game and utilizing the 5.1 surround sound capability to give the viewer a sense of immersion and of "being there." These objectives are achieved by use of two separate microphone setups common to UK Premier League coverage and also to coverage of World Cup and other international football. For on-pitch sounds the ball-kicks and whistle-blows are happening some distance from any possible microphone position so shotgun microphones are used to pick them up. Twelve shotgun microphones are positioned around the football pitch facing towards the action. If all of the microphones are live in the mix at any given time the background noise from the crowd swamps the sounds from the pitch making them inaudible. In order to prevent this from happening, microphones are mixed live so that only the microphone, or microphones, closest to the ball location is in the mix at any given time[1].

---

[1] Although this setup is common for football, capture techniques vary between live events.

This requires a skilled sound engineer to follow the action on the pitch on the mixing console in the outside broadcast truck and ensure that only the appropriate microphones are active in the mix. As the broadcast is live the engineer must predict where the ball is going to be next but also has to be aware of what likely camera angles will be chosen by the producer. At any given moment during the event between one and three microphones will be active in the mix. All of these microphones are panned centrally, either to a central loudspeaker, or more often to a phantom center between left and right in order to avoid any potential issues from the downmixing process. The crowd sound for live football coverage is considered key to building the atmosphere for the television viewer and is usually picked up by a single SoundField microphone suspended from the gantry above the crowd. For television broadcast the Z (height) component of the B-format mic encoder output is ignored and the B-format signals are decoded into a 5.1 feed at the OB truck. This 5.1 crowd noise channel is mixed into surround and left and right channels of the 5.1 program audio both to give a more immersive experience for viewers and also to cover up any audible artifacts from mixing between the pitch-side microphones. Although the pitch-side shotgun microphones pick up many of the sounds on the pitch that are of interest the audio feeds from these also contain large amounts of crowd noise. Trying to separate these on-pitch sounds by reducing the mix of the SoundField microphone dedicated for crowd noise leads to unpleasant and disorientating effects as microphones in different areas of the stadium are faded in and out. Therefore in order to provide an object-based solution accessible to hard of hearing viewers such as that described there was a need to develop a method of separating out on-pitch sounds effectively from crowd noise.

3.3.1 Method: New Production Techniques for Audio Acquisition

In order to extract on-pitch sounds from the audio scene as audio objects, techniques were developed that were designed to cause minimum change to the current workflows of sound engineers. The methodologies adopted utilize the same microphones that are already used and were designed to provide a largely automated system for the broadcast production team. Considerable research was carried out into existing workflows, interviews were held with sound engineers from SISLive and Sky and site visits carried out to outside broadcasts to elicit a detailed understanding of the processes currently carried out and of the problems associated with providing robust separation of sounds that could both be used for spatially positioning the resultant audio object and that would be potentially useful for comprehension and enjoyment of hard of hearing people in more traditional broadcast scenarios. Microphone feeds from every microphone used were captured on-site including all pitch-side microphones and a SoundField microphone and an Eigenmike [65] capturing crowd noise. These were stored on hard disc for later processing along with a separate BBC commentary feed. Audio object templates were developed

for each class of sound that was required to be captured as a discrete object; in this case templates for ball-kicks and whistle-blows based on envelope and spectral content were created. Software was developed that monitored every pitch-side microphone, comparing it to the stored audio object template. When a matching sound event was detected in a microphone feed all other microphone feeds were scanned for matching characteristics to identify all microphones that had some element of the event present in its feed. For every pair of microphone feeds that picked up the event a hyperbola, along which the event must have occurred, was calculated based on the difference in time of arrival of the sound at the two microphones. Where more than two microphone pairs had captured the sound event the intersections of the calculated hyperbolas gave an accurate coordinate location for the sound event. In the FascinatE project, which utilized object-based audio for spatial positioning of audio objects in surround and 3D reproduction systems, these coordinate locations were used to automatically position the sound dependent on a user-defined viewpoint, by a virtual director system [66] or on a viewpoint defined by the production team. In the case of audio objects for clean audio the critical factor is different; the audio objects are panned centrally, as in current broadcast, and the key outcome is to identify the audio event and extract it from the acoustic scene in isolation from crowd noise and other sounds that may be detrimental to clarity, thus enabling freedom to mix the relative levels of the individual components at the user-end. In order to accomplish this, the microphone feed containing the loudest example of the sound event was identified based on the derived location of the sound event. Once this microphone was identified an envelope was applied to the microphone feed based on the temporal characteristics of the detected sound event. In this way relevant pitch-side microphones were only ever active in the produced mix for the duration of the sound event. The short section of crowd noise picked up by that microphone was then effectively masked by the sound event itself. A flow diagram for the object extraction process can be seen in Fig. 3. The resultant audio object, together with its paired coordinate metadata was coded into the broadcast audio stream for decoding at the rendering device.

The user of the system was presented with an interface enabling selection of pre-determined reproduction mixes the input of which was three discrete streams of audio:

1) Clean BBC commentary feed with no background noise taken directly from the commentary microphone;
2) Crowd noise from an Eigenmike located on the gantry at the camera position;
3) On-pitch sounds extracted using the audio object extraction techniques described here.

The three streams were combined and transmitted over IP to the audio part of the FascinatE Render Node (FRN). The FRN consisted of the Audio Composer (developed by University of Salford) and the Audio Presenter (developed by Technicolor). The Audio Composer was responsible for
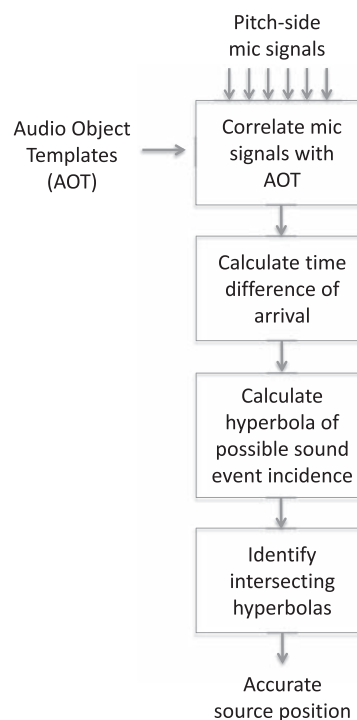

Fig. 3. Flow diagram illustrating audio object extraction developed in the FascinatE project

decoding the various audio streams and metadata received, arranging them spatially and mixing them based on user input and on scripts generated as part of the production. User input included the chosen user-defined scene incorporating the pan, tilt, and zoom information necessary to spatially position the on-pitch audio appropriately for the chosen camera view. Additionally user-choices of the relative levels of foreground (commentary and on-pitch) and background (crowd) sound could also be selected. Script information was generated in two ways, at both production and user end.

At the production end, scripting data was generated both automatically and manually including automatically chosen regions of interest determined by the Virtual Director (developed for FascinatE by the Joanneum Research Institute) or by producers. At the user-end user preferences in the rendering device would provide further scripting input included choices based on user access requirements such as preferred speech level relative to background noise. Such an object-based approach could therefore allow bespoke processing to be done on selected audio elements to suit a given person's preferences in terms of their individual hearing loss, i.e., various multiband compression and filtering schemes could be applied to the speech only, in a manner similar to the techniques used in hearing aids. Further to this the user's own reproduction system/setup could be used to perform a basic audiogram that would allow the individualized processing parameters to be defined and fed in to the scripting engine. It would hence be possible to perform processing on individual elements of the audio scene to result in the best possible listener experience.
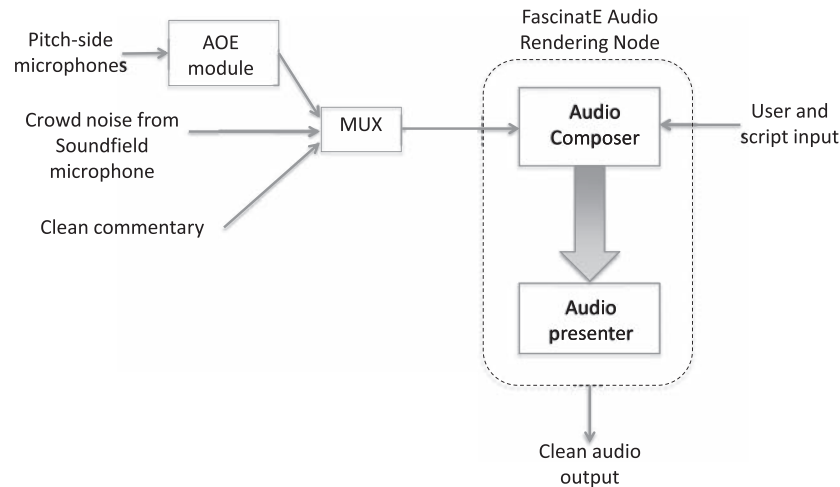
Fig. 4. Clean Audio in FascinatE: Extracted audio objects, crowd noise, and clean commentary multiplexed for transmission, decoded, and reproduced by the FascinatE audio render node

4 CONCLUSIONS

An object-based approach to clean audio, combined with methods to isolate sounds that are important to the narrative and meaning of a broadcast, has the potential to enable users to have complete control of the relative levels of all aspects of audio from TV broadcast. Any of the solutions previously discussed, such as non-speech channel attenuation, dynamic range control, and other processing could then be carried out on only the speech, or other important content of the program, in isolation depending on user preferences and on the nature and genre of the program material. Although the FascinatE implementation described here was for live event broadcast, object-based methods could be applied to any genre—live sports probably being one of the most challenging because of the difficulties in extracting the audio objects. The main limitation of the system described was that of latency. In the FascinatE project there was sufficient video latency that all processing could take place and the resultant stream of audio objects, together with positional metadata, could be streamed alongside the panoramic video stream to the end user. There are substantial challenges associated with adapting the technique for real-time implementation in today's broadcast systems, namely a much reduced video processing latency of around 20 ms. This technique for provision of clean audio was demonstrated using recorded microphone feeds at the FascinatE project's final demonstration at the University of Salford's campus at MediaCityUK on 30 May 2013. During the demonstration visitors were able to alter the balance between foreground and background audio where the foreground audio consisted of commentary and on-pitch sounds and the background consisted of crowd noise.

5 ACKNOWLEDGMENT

6 REFERENCES

[1] Action on Hearing Loss, "Hearing Loss Statistics" (2014). URL http://www.actiononhearingloss.org.uk/your-hearing/about-deafness-and-hearing-loss/statistics.aspx, accessed 5 December 2014.

[2] Voice of the Listener & Viewer, "VLV's Audibility of Speech on Television Project Will Make a Real Difference," VLV News Release, March 2011. Available at http://www.vlv.org.uk/documents/06.11Pressreleasefrom VLV-AudibilityProject-0800hrs1532011_002.pdf, accessed 5 December 2014.

[3] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, "MPEG-H Audio—The New Standard for Universal Spatial/3D Audio Coding," presented at the *137th Convention of the Audio Engineering Society* (2014 Oct.), convention paper 9095.

[4] S. Meltzer, M. Neuendorf, D. Sen, and P. Jax, MPEG-H 3D Audio—The Next Generation Audio System (Institution of Engineering and Technology, 2014).

[5] MPEG, "ISO/IEC 23008, High Efficiency Coding and Media Delivery in Heterogeneous Environments (open)" (2013).

[6] A. C. Davis, "The Prevalence of Hearing Impairment and Reported Hearing Disability among Adults in Great Britain," *Int'l J. Epidemiology*, vol. 18, no. 4, pp. 911–917 (1989).

[7] W. Noble, G. Naylor, N. Bhullar, and M. A. Akeroyd, "Self-Assessed Hearing Abilities in Middle- and Older-Age Adults: A Stratified Sampling Approach," *Int'l J. Audiology*, vol. 51, no. 3, pp. 174–180 (2012).

[8] K. Palmer, M. Griffin, H. Syddall, A. Davis, B. Pannett, and D. Coggon, "Occupational Exposure to Noise and the Attributable Burden of Hearing Difficulties in Great

Britain," *Occup. Environ. Med.*, vol. 59, no. 9, pp. 634–639 (2002).

[9] K. T. Palmer, D. Coggon, H. E. Sydall, B. Pannett, and M. J. Griffin, "CRR 361/2001 Occupational Exposure to Noise and Hearing Difficulties in Great Britain," Technical Report, University of Southampton (2001).

[10] Health and Security Executive (HSE), "Health and Safety Statistics 1996/7," Technical Report, Health and Safety Commission (1997).

[11] Health and Security Executive (HSE), "Statistics: Noise Induced Deafness Summary," available at http://www.hse.gov.uk/noise/statistics.htm, accessed: 5 December 2014.

[12] W. W. Clark, "Noise Exposure from Leisure Activities: A Review," *J. Acous. Soc. Amer.*, vol. 90, no. 1, pp. 175–181 (1991).

[13] M. Maassen, W. Babisch, K. D. Bachmann, H. Ising, G. Lehnert, P. Plath, P. Plinkert, E. Rebentisch, G. Schuschke, M. Spreng, G. Stange, V. Struwe, and H. P. Zenner, "Ear Damage Caused by Leisure Noise," *Noise & Health*, vol. 4, no. 13, pp. 1–16 (2001).

[14] A. Axelsson, A. Eliasson, and B. Israelsson, "Hearing in Pop/Rock Musicians: A Follow-Up Study," *Ear and Hearing*, vol. 16, no. 3, pp. 245–253 (1995).

[15] A. Axelsson and F. Lindgren, "Hearing in Pop Musicians," *Acta Oto-laryngologica*, vol. 85, nos. 1-6, pp. 225–231 (1978).

[16] S. Sadhra, C. A. Jackson, T. Ryder, and M. J. Brown, "Noise Exposure and Hearing Loss among Student Employees Working in University Entertainment Venues," *Annals of Occupational Hygiene*, vol. 46, no. 5, pp. 455–463 (2002).

[17] J.-H. Peng, Z.-Z. Tao, and Z.-W. Huang, "Risk of Damage to Hearing from Personal Listening Devices in Young Adults," *J. Otolaryngology*, vol. 36, no. 3, pp. 179–183 (2007).

[18] BBC, "Clear Sound: Best Practice Tips" (2011), [cited 14 August 2013]; BBC Academy training guide. Available from: http://www.bbc.co.uk/academy/production/article/art20130702112135255, accessed 5 December 2014.

[19] Royal National Institute for Deaf People (RNID), Annual Survey Report 2008 (2008).

[20] Royal National Institute for Deaf People (RNID), Annual Survey Report 2005 (2005).

[21] ITU-R 54/6, Sound Systems for the Hearing Impaired (1994).

[22] C. D. Mathers, "A Study of Sound Balances for the Hard of Hearing," Technical Report, BBC (1991).

[23] D. J. Meares, "R&D Report 1991–14: HDTV Sound: Programme Production Developments," Technical Report, BBC (1991).

[24] W. Hoeg and T. Lauterbach, *Digital Audio Broadcasting: Principles and Applications of DAB, DAB+ and DMB* (Wiley, 2009).

[25] C. Uhle, O. Hellmuth, and J. Weigel, "Speech Enhancement of Movie Sound," presented at the *125th Convention of the Audio Engineering Society* (2008 Oct.), convention paper 7628.

[26] M. Armstrong, "Audio Processing and Speech Intelligibility: A Literature Review," Technical Report, BBC (2011).

[27] P. Kendrick and B. Shirley, "Performance of Independent Component Analysis When Used to Separate Competing Acoustic Sources in Anechoic and Reverberant Conditions," presented at the *124th Convention of the Audio Engineering Society* (2008 May), convention paper 7413.

[28] A. R. Carmichael. "Evaluating Digital 'On-Line' Background Noise Suppression: Clarifying Television Dialogue for Older, Hard-of-Hearing Viewers," *Neurophysical Rehabilitation*, vol. 14, nos. 1-2, pp. 241–249 (2004).

[29] H. Müsch, "Aging and Sound Perception: Desirable Characteristics of Entertainment Audio for the Elderly," presented at the *125th Convention of the Audio Engineering Society* (2008 Oct.), convention paper 7627.

[30] F. Rumsey, "Hearing Enhancement," *J. Audio Eng. Soc.*, vol. 57, pp. 353–359 (2009 May).

[31] G. Cervellera and A. Quaranta, "Audiologic Findings in Presbycusis," *J. Auditory Res.*, vol. 22, no. 3, pp. 161–171 (1982 Jul.).

[32] E. L. Stine, A. Wingfield, and L. W. Poon, "How Much and How Fast: Rapid Processing of Spoken Language in Later Adulthood," *Psychology and Aging*, vol. 1, no. 4, p. 303 (1986).

[33] P. Rabbitt, "Mild Hearing Loss Can Cause Apparent Memory Failures which Increase with Age and Reduce with IQ," *Acta Oto-laryngologica*, vol. 111, no. S476, pp. 167–176 (1991).

[34] EBU, "EBU – TECH 3333: EBU HDTV Receiver Requirements," Geneva, Switzerland (2009 Mar.).

[35] Open IPTV Forum, "OIPF Release 2 Specification Volume 2 – Media Formats" (2011 Jun.).

[36] 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).

[37] UK Clean Audio Forum, "Liaison Statement from UK Clean Audio Forum to ITU FG IPTV," Liaison Statement, International Telcommunications Union Focus Group on IPTV (2007 Jan.).

[38] R. Dressler, "A Step Toward Improved Surround Sound Making the 5.1-Channel Format a Reality," presented at the *100th Convention of the Audio Engineering Society* (1996 May), convention paper 4287.

[39] T. Holman, "The Number of Loudspeaker Channels," *Proc. 19th International Conference of the Audio Engineering Society: Surround Sound-Techniques, Technology, and Perception* (2001 Jun.), conference paper 1906.

[40] T. Holman, "The Number of Audio Channels," presented at the *100th Convention of the Audio Engineering Society* (1996 May), convention paper 4292.

[41] B. Shirley, P. Kendrick, and C. Churchill, "The Effect of Stereo Crosstalk on Intelligibility: Comparison of a Phantom Stereo Image and a Central Loudspeaker Source," *J. Audio Eng. Soc.*, vol. 55, pp. 852–863 (2007 Oct.).

[42] D. Clark, "Measuring Audible Effects of Time Delays in Listening Rooms," presented at the *74th Convention*

of the Audio Engineering Society (1983 Oct.), convention paper 2012.

[43] D. H. Cooper and J. L. Bauck," "Prospects for Transaural Recording," *J. Audio Eng. Soc.*, vol. 37, pp. 3–19 (1989 Jan./Feb.).

[44] J. L. Bauck and D. H. Cooper, "Generalized Transaural Stereo," presented at the *93rd Convention of the Audio Engineering Society* (1992 Oct.), convention paper 3401.

[45] E. Vickers, "Fixing the Phantom Center: Diffusing Acoustical Crosstalk," presented at the *127th Convention of the Audio Engineering Society* (2009 Oct.), convention paper 7916.

[46] R. Bücklein, "The Audibility of Frequency Response Irregularities," *J. Audio Eng. Soc.*, vol. 29, pp. 126–131 (1981 Mar.).

[47] G. S. Kendall, "The Decorrelation of Audio Signals and its Impact on Spatial Imagery," *Computer Music J.*, vol. 19, no. 4, pp. 71–87 (1995).

[48] M. Boueri and C. Kyriakakis, "Audio Signal Decorrelation Based on a Critical Band Approach," presented at the *117th Convention of the Audio Engineering Society* (2004 Oct.), convention paper 6291.

[49] G. Augspurger, S. Bech, R. Brook, E. Cohen, J. Eargle, and T. A. Schindler, "Use of Stereo Synthesis to Reduce Subjective/Objective Interference Effects: The Perception of Comb Filtering, Part II," presented at the *87th Convention of the Audio Engineering Society* (1989 Oct.), convention paper 2862.

[50] 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 (2009).

[51] A. Carmichael, H. Petrie, F. Hamilton, and J. Freeman, "The Vista Project: Broadening Access to Digital TV Electronic Programme Guides." *Psych. J.*, vol. 1, no. 3, pp. 229–241 (2003).

[52] M. M. Goodwin and J.-M. Jot, "Primary-Ambient Signal Decomposition and Vector-Based Localization for Spatial Audio Coding and Enhancement," *Proc. Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, vol. 1, pages 1–9 (2007).

[53] S. Zielinski, J. Mayne, and F. Rumsey, "Improving Intelligibility of Surround Sound Using Principal Component Decomposition," *Proc. Institute of Acoustics, 21st Reproduced Sound Conference*, vol. 27, pt. 5 (2005).

[54] B. Shirley, "*Improving Television Sound for People with Hearing Impairments*," Ph.D. Thesis (2013).

[55] C. Q. Robinson, S. Mehta, and N. Tsingos, "Scalable Format and Tools to Extend the Possibilities of Cinema Audio," *SMPTE Motion Imaging J.*, vol. 121, no. 8, pp. 63–69 (2012).

[56] A. D. Kraemer, J. Tracey, and T. Katsianos, US patent application 13/791,488: Object-oriented audio streaming system (2013).

[57] O. Schreer, J.-F. Macq, O. A. Niamut, J. Ruiz-Hidalgo, B. Shirley, G. Thallinger, and G. Thomas *Media Production, Delivery and Interaction for Platform Independent Systems: Format-Agnostic Media* (John Wiley & Sons, 2013).

[58] H. Fuchs, S. Tuff, and C. Bustad, "Dialogue Enhancement—Technology and Experiments," Technical Report, EBU (2012).

[59] H. Fuchs and D. Oetting, "Advanced Clean Audio Solution: Dialogue Enhancement," *SMPTE Motion Imaging J.*, vol. 123, no. 5, pp. 23–27 (2014).

[60] Joanneum Research, Technicolor, TNO, University of Salford, BBC, TNO, Softeco Sismat, Interactive Institute, Fraunhofer HHI, ARRI, and Universitat de Catalunya. Fascinate project [FP7/2007-2013] under grant agreement no. [248138] (2010).

[61] P. Kauff, P. Eisert, J. Schuessler, C. Weissig, and A. Finn, US patent application 14/049,057: Capturing panoramic or semi-panoramic 3D scenes (2014).

[62] I. Forrester and A. Churnside, "The Creation of a Perceptive Audio Drama," *Proc. NEM Summit* (2012).

[63] S. E. Carlsson "Diegetic and Non-Diegetic Sounds," [cited 2013 September]; available from: http://filmsound.org/terminology/diegetic.htm, accessed 5 December 2014.

[64] R. G. Oldfield, B. Shirley, and J. Spille, "Object Based Audio for Interactive Football Broadcast," *Multimedia Tools & Applications*, DOI:10.1007/s11042-013-1472-2 (2012).

[65] C. Barnhill, J. West, and I. Busch-Vishniac, "Impulse Response of a Spherical Microphone Array (Eigenmike)," *J. Acous. Soc. Amer.*, vol. 119, no. 5, pp. 3378 (2006).

[66] EBU, "EBU HDTV Receiver Requirements" (2009 Mar.).

THE AUTHORS

Ben Shirley   Rob Oldfield

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. Dr. Shirley was principal investigator on the ITC and Ofcom funded Clean Audio project that investigated methods to improve TV sound for people with hearing impairments and was a member of the UK Clean Audio Forum that contributed to ETSI and EBU digital video broadcast standards. He led the VoIPText project developing text interfaces for people with hearing impairments. He has been a consultant on a wide range of audio related commercial projects. More recently he was Principal Investigator for University of Salford's work in the EU FP7 FascinatE project that developed an end-to-end interactive future broadcast system. He is currently working on the EPSRC funded S3A project investigating future object-based spatial audio systems. Although primarily working in science and technology he also enjoys the challenge of cross-disciplinary projects and has worked with composers, musicians, film makers, choreographers, and artists in projects spanning areas where technology and creativity overlap. Dr. Shirley's teaching includes audio, broadcast, and music technology and audio production/post production at the University of Salford's MediaCityUK campus and he has been a visiting guest lecturer delivering masterclasses on sound design and 3D audio at Carnegie Mellon's Entertainment Technology Center.

●

Dr. Rob Oldfield is an Audio Research Consultant at the University of Salford. Rob received his Masters by research at the University of Salford in 2006 for his thesis on tunable membrane absorbers for room acoustic applications; staying at the same institution Rob went on to receive his PhD in 2013. His doctoral thesis was on the analysis and improvement in the rendering of focused source reproduction using wave field synthesis. Rob's main research interests are in spatial audio, room acoustics, broadcast audio and signal processing and he has been a member of the Acoustic Research Centre at Salford University for over 10 years. Dr Oldfield has worked on many research projects during that time at the University including projects for Defra, Ofcom and being co-investigator on the EU FP7 funded FascinatE project. Rob's current work as an audio research consultant provides opportunities to work on many diverse research and consultancy projects in the audio and acoustics fields as part of work funded by both commercial entities and also research councils.