L. A. Ward and B. G. Shirley, "Personalization in Object-Based Audio for Accessibility: A Review of Advancements for Hearing Impaired Listeners"
J. Audio Eng. Soc., vol. 67, no. 7/8, pp. 584–597, (2019 July/August.).
DOI: https://doi.org/10.17743/jaes.2019.0021

Personalization in Object-Based Audio for Accessibility: A Review of Advancements for Hearing Impaired Listeners

LAUREN A. WARD, *AES Student Member,* **AND BEN G. SHIRLEY,** *AES Member*
(L.Ward7@edu.salford.ac.uk) (B.G.Shirley@salford.ac.uk)

Acoustics Research Centre, University of Salford, Manchester, UK

Hearing loss is widespread and significantly impacts an individual's ability to engage with broadcast media. Access can be improved through new object-based audio personalization methods. Utilizing the literature on hearing loss and intelligibility this paper develops three dimensions that are evidenced to improve intelligibility: spatial separation, speech to noise ratio, and redundancy. These can be personalized, individually or concurrently, using object-based audio. A systematic review of all work in object-based audio personalization is then undertaken. These dimensions are utilized to evaluate each project's approach to personalization, identifying successful approaches, commercial challenges, and the next steps required to ensure continuing improvements to broadcast audio for hard of hearing individuals.

0 INTRODUCTION

Hearing loss affects 16% of the United Kingdom's population [1], with similar statistics reflected throughout Europe and North America [2, 3]. An aging demographic and the prevalence of age-related hearing loss suggests that this statistic is likely to rise [4, 5]. People with some degree of hearing loss are therefore making up an increasing percentage of television audiences [6]. Furthermore, those over 50 years old in the United States of America and those over 55 in the United Kingdom watch more television on average than any other age demographic in their respective countries [7, 8]. However, those with hearing loss often have difficulty in understanding broadcast media [9, 10, 6, 11]. Recent developments in broadcast technology have the potential to deliver real accessibility improvements for hard of hearing people. In particular the roll-out of next-generation object-based audio (OBA) formats has the capability to allow audiences to personalize aspects of the content to their needs [6, 12–15].

This paper outlines the current barriers to broadcast access faced by hard of hearing individuals. Three dimensions of audio personalization are derived from the literature on intelligibility and hearing loss. These are used to systematically review the current work in OBA personalization. Finally challenges and proposed future directions are discussed.

1 BROADCAST AND HEARING IMPAIRMENT

Increasingly, human communication and the dissemination of news and information is achieved through audiovisual content. This extends beyond terrestrial broadcast to the internet, with audiovisual content making up 73% of all internet traffic in 2016 and projected to increased to 82% by 2021 [16]. The importance of audiovisual media for education, entertainment, and national identity is recognized by international standards bodies such as the ITU [17] as well as legislatively in the charters of numerous national broadcasters [18, 19]. The UN Convention on the Rights of Persons with Disabilities Articles 9 and 21 emphasize this right of access to information, communication, and mass media services for those with disabilities [20].

Beyond this, access to television is valued by hearing impaired individuals and can provide vital social inclusion. Eighty-four percent of the hearing-impaired participants in a recent survey reported that hearing well when watching TV/video was "very important" or "extremely important" [11]. Coupled with increased rates of depression and social isolation among adults with even mild to moderate hearing loss [21], it is a social imperative to provide the requisite broadcast accessibility services for those with hearing loss.

The effective design of such accessibility strategies requires a comprehensive understanding of the characteristics, prevalence, and challenges of hearing loss. The

remainder of this section summarizes the characteristics of hearing loss and its effect on broadcast accessibility.

1.1 Barriers to Accessing Broadcast Content

Headlines decrying inaudible television dialog have become common of late [22, 23] and the issue was deemed significant enough to be debated in the UK Parliament's Upper House [24]. However, this problem is not a new one and research into dialog clarity has been ongoing for more than 25 years [25]. A large-scale 2008 study by the Royal National Institute for the Deaf reported that 87% of hard of hearing viewers struggled with television speech [9]. Similar difficulty was reflected in a cross-sectional survey of a single evening's viewing carried out by the BBC [10]. This showed 60% of viewers had difficulty hearing the speech in the broadcasts at some point during the evening. It also identified 4 main factors affecting speech understanding: clarity of speech, unfamiliar or strong accents, background noise, and background music [26].

While identifying such problems are quite straightforward, locating their origin in the broadcast chain and mitigating them is more complex. Armstrong explores this complexity, defining a problem space covering the numerous points where audibility can be degraded; from the original performance and content capture, through production and broadcast, to reproduction in the home [6]. Mapp defines a similar space, including three specific listener-based factors: hearing acuity, attention and alertness, and familiarity/fluency of language [27]. Armstrong and Crabb consider these human factors in terms of an individuals' "media access needs," of which two types are defined: sensory and cognitive [12]. Sensory needs impact on an individual's ability to perceive broadcast content and may be permanent (e.g., hearing loss) or temporary (e.g., consuming content in a noisy environment). Cognitive accessibility refers to an individual's ability to understand, engage with, and enjoy content. Meeting cognitive needs is about processing and comprehending the information (given differing language, cultural knowledge, and norms) and memory of what has occurred within the program. This review focuses on the sensory needs of those with hearing loss.

1.2 Prevalence of Hearing Impairment

In 2015 the charity *Action on Hearing Loss* (AoHL) estimated 11 million people in the United Kingdom were affected by hearing loss [1]. These statistics are mirrored in countries with similar demographics such as Australia [28] (2006) and the United States [2] (2003–2004). The WHO estimate that over 360 million people worldwide have a disabling degree of hearing loss [29]. AoHL project that by 2035, the number of individuals with hearing loss in the UK will 15.6 million people [1], which is in part due to an aging population [4]. Presbycusis, age-related hearing loss [30], is the single largest cause of hearing loss in the UK [1]. Another major cause is noise-induced hearing loss, often from occupational exposure [31, 32], though this can also be from recreational activities such as concerts [33].

1.3 Characterization of Hearing Loss

Hearing loss is often characterized by the location of the impairment within the auditory system: *conductive* loss is as a result of problems within the ear canal, ear drum or middle ear; *sensorineural* loss is as a result of problems with the inner ear; and *mixed* loss combines these factors [34]. Presbycusis and noise-induced hearing loss are common types of sensorineural loss [35]. The severity of hearing loss is often characterized by pure-tone threshold audiometry, grouping individuals into four categories: mild, moderate, severe, and profound [36, 1]. Those with mild loss (in the range 20 – 40 dB) generally struggle to understand speech in noisy situations but often are able to understand speech in quiet unaided [1, 37, 38]. Those with moderate loss (41 − 70 dB) generally have difficulty understanding speech under any condition without a hearing aid [1]. Those with mild to moderate are the majority of hearing impaired individuals in the UK (91.7%). Those with severe (71 − 95 dB) to profound loss (>95dB) often rely on lip-reading, powerful hearing aids or cochlear implants. However, generally only a small proportion of those who could benefit from a hearing aid actually have one fitted (24% in Australia [39]), and many of those who have had a hearing aid fitted do not use them regularly [40, 39].

Pure-tone threshold audiometry provides a common and readily understood definition of hearing loss severity. However, audiometric thresholds do not fully account for the variability in individuals' ability to understand speech in noise [41–43]. Even listeners with normal thresholds can vary significantly in their ability to understand speech in noise [44, 45]. For this reason, evaluations of speech in noise performance are becoming increasingly used in clinical settings [41], usually quantifying the point where the listener can understand 50% of the speech [46–48].

The difference between audiometric thresholds and speech in noise performance was modeled by Plomp in 1978 [43], who defined two classes of hearing loss:

- *Attenuation*: loss due to the reduction in perceived volume as a result of reduced audiometric thresholds;
- *Distortion*: comparable to a decreased effective speech to noise ratio. Also termed suprathreshold.

Contributors to suprathreshold loss include reduction in sensitivity to temporal fine structure [49], loudness recruitment (reduced dynamic range) [50], and reduction in frequency resolution [51]. Some of these factors are collectively termed "hidden hearing loss," hypothesized to be from cochlear neuropathy (loss of the high frequency auditory nerve fibers) [52, 53]. Non-audiometric components of hearing have also been shown to decline with age [54, 45, 55], possibly due to age-related loss of acuity of subcortical neural temporal coding of sound [56, 57].

Fully characterizing hearing loss is an ongoing research challenge. As such, the dilemma facing those developing accessibility services has echoes of the *Anna Karenina* principle (with deference to Leo Tolstoy): "*normal hearing listeners are all alike; every hearing impaired listener is*

hearing impaired in their own way" [42, 58]. A personalized approach therefore has the greatest chance of successfully creating accessible audio; however mass media present a challenging dichotomy between audience-wide and individual needs.

2 METHODOLOGY OF THE SYSTEMATIC REVIEW

No previous systematic reviews in this area exist, with the most comprehensive review paper being a 2016 BBC whitepaper [6]. The paper concludes that successful delivery of accessible audio is technically possible through personalized object-based audio (OBA) content but broad uptake requires better understanding of user wants and needs. For this reason, this paper revisits the topic to evaluate the progress made and systematically collate evidence for different approaches. The results of this aim is to support technology and content creators and broadcasters in prudent decision making around OBA implementation.

This aim is achieved with a two stage review methodology. Stage one performs a conceptual analysis of broadcast speech intelligibility literature and pre-OBA research, to establish the existing evidence for different personalization approaches (Sec. 3). Three main types were identified; speech to noise ratio, spatial separation, and redundancy. They are termed here as dimensions of personalization given that these approaches can coexist to varying degrees. The second stage utilized a systematic review methodology, summarizing all active and completed projects on OBA personalization (Sec. 4). These were then analyzed with reference to the three personalization dimensions (Sec. 5).

This review considers the literature in terms of projects due to the evolving and active nature of the research area, similarly to [6]. Projects are defined here as an individual or collaborative investigations with a specified aim, supported by one or more publications (including but not limited to peer-reviewed literature and public project deliverables). Only research publicly available prior to July 12, 2018 was considered. For inclusion a project aims had to meet the following criteria:

- To enable personalization of an element of broadcast audio, which the end-user has control of at time of consumption; AND
- Use OBA to do so; OR
- Rely on OBA methods for eventual implementation of a theoretical investigation.

Sixteen projects meeting this criteria were identified. Two additional projects were initially identified and later excluded as they did not specifically describe *audio* personalization [59, 60]. The majority of projects addressed speech to noise ratio (15 projects). Only 3 projects explored personalization of spatial separation and redundancy respectively.

3 DIMENSIONS OF PERSONALIZATION

This section outlines the dimensions of personalization identified through a conceptual analysis of broadcast speech intelligibility literature. Speech intelligibility is often defined as the proportion of words that are correctly identified, distinguishing it from comprehension [61]. It may also be defined as the proportion of words understood [62], incorporating elements of comprehension and quality. The latter definition is used here.

Furthermore, the definition used considers two types of intelligibility. Signal-dependent intelligibility is based solely on the availability of the speech signal. Complementary intelligibility however utilizes other, non-speech, cues from the speech signal. These can include syntax, semantics, and multi-modal cues such as facial expressions [63]. These complementary cues are also referred to as "top-down information" [64], and they have been shown to play an increased role in speech perception when hearing is challenged, either by hearing impairment, or by masking from competing sources [65, 66]. It has been proposed that complementary intelligibility is a result of the manner in which the brain composes perceptual auditory objects using expectations to predict unavailable parts of the input signal [67].

3.1 Speech to Noise Ratio

The first response when speech is not understood on television is to turn up the volume [68, 11]. However it is commonly reported by hard of hearing listeners that despite having the television at near full volume, it does not aid in following on screen conversations [69]. When the *attenuation* caused by hearing loss is overcome, the *distortion* loss, or effective speech to noise reduction, remains [43].

Two main studies have investigated the adjustment of speech to noise ratio to improve intelligibility. The first was conducted by Mathers in 1991 with the BBC and other partners [25]. This used audiovisual clips with either +6 dB, –6 dB or unchanged background sound levels and subjective ratings of quality were elicited from participants. This suggested that a 6 dB reduction to the background sounds produces a small improvement in quality but the study lacked statistical analysis of its findings. The second, more recent study was conducted in 2010—the BBC Vision Audibility project. It used a similar experiment to Mathers, providing three mixes with varying background sound levels to participants: +4 dB, –4 dB, and unchanged [6]. This showed that greater levels of background sound definitely inhibited speech understanding but less background sound did not always provide an improvement.

The inability of these studies to determine a single, optimal speech to noise ratio is unsurprising given that speech reception threshold is itself a feature used to characterize the degree of an individual's disablement from hearing loss. Research conducted into speech enhancement for television has shown that this variability is further complicated when the speech and background cannot be controlled separately and post hoc speech enhancement is used [70, 71]. As such its application to improving broadcast access has shown

limited efficacy [72]. As a personalizable parameter, however, it has significant potential for improving accessibility for hard of hearing listeners.

3.2 Spatial Separation

For normal hearing listeners, speech intelligibility improves when the masker and target speech are spatially separated [73]. Hearing impaired listeners can also benefit from spatial release from this masking, though to a reduced degree [74], dependent on their specific hearing impairment and localization ability. Part of the Clean Audio project explored this, evaluating the effect of reproduction using a phantom center compared with a central loudspeaker on speech intelligibility [75, 76]. This showed a measurable improvement in intelligibility of up to 5.9% when using a central loudspeaker. If speech is placed in the center channel, this can effectively improve intelligibility; however, the placement of speech, and consequent efficacy of this approach, is dictated by the preferences of producers and broadcasters. OBA gives the potential for the spatial location of the speech to be a personalizable dimension, through the separate control of an object's location.

3.3 Redundancy

Human listeners use a variety of tactics to hear speech, even buried in noise, which in film we call music and effects. [77]

Redundancy, or additional information that may be superfluous for normal hearing listeners in quiet, can facilitate understanding in less favorable conditions or for people with hearing loss. This is particularly relevant to audio visual media. Audiovisual content does not represent a standard speech in noise problem [78, 79]. Non-speech broadcast content includes music, effects, Foley, and ambiences as well as noise; therefore the speech in noise problem faced by hard of hearing viewers is not as simple as having a target (speech) and masker (noise). Non-speech signals can provide redundancy and improve complementary intelligibility. These cues can come from within the speech and other audio signals (single mode), or may come from other sources such as accompanying visuals (multi-modal). The most commonly used accessibility service for hard of hearing people is captioning (also known as subtitles) which, for people with some residual hearing, provides redundant information. In a 2015 study into subtitle usage, one subtitle user described the role of subtitles for them as: *"... so I'm reading and hearing but the hearing only works if I'm reading— putting two and two together"* [80].

The importance of redundancy in understanding speech has long been understood in terms of context, word familiarity, and syntactic structure probability [81, 82, 64]. Research by Bilger in 1984 showed that word recognition in noise by older listeners with sensorineural hearing loss more than doubled when the speech was semantically predictable (from recognizing 37% of keywords up to 76%) [82]. Recent adaptations of his work consistently demonstrate this effect [83–86].

One redundancy cue shown to provide improvements in intelligibility is familiarity with the speaker. A study by Souza et al. found that in both noise and quiet, hard of hearing listeners could understand speech in noise better when spoken by a familiar voice, spouse or close friend, than by a stranger [87]. Even for speech that is previously unfamiliar, familiarizing a listener with the speaker's voice beforehand can result in intelligibility gains [88].

Other types of single mode cues can be provided by non-speech sounds, an area that controlled studies have only recently explored [89–91]. A 2016 study by Hodoshima showed that some types of preceding sounds aid intelligibility of urgent public address style speech [89]. A 2017 study by Ward et al. showed that the inclusion of sound effects related to the keywords can improve word recognition rates in noise from 36% to 61% for normal hearing listeners. Follow-on work demonstrated that the same effect is present for some hard of hearing listeners, although the effect is dependent on hearing acuity in the better hearing ear. Those with mild loss exhibited comparable benefits to normal hearing listeners [91].

Multi-modal cues have been investigated in many studies, including the interaction of different complementary intelligibility cues [92–94]. In Augert et al.'s work the effect of prosody and pictorial situational context was investigated with young (5–9 years old) French speaking listeners [92]. It showed that by age five children can utilize the situational context of speech [92]. Zekveld et al. analyzed the effect of semantic context and related and unrelated text cues on speech intelligibility, showing that both relevant and irrelevant semantic context influences speech perception in noise [93]. Spehar et al. has investigated the effects of different types of contextual cues showing that participants benefited from both visual and speech-based context [94]. Multi-modal redundancy is well illustrated by findings from the Clean Audio Project. Using a forced choice comparison test between video clips with hearing impaired participants, results indicated a statistically significant correlation between video clip preference and the percentage of face-to-camera dialog for both speech clarity and enjoyment ratings [78]. Interestingly, participants were overall unaware that they were lip-reading.

There are a multitude of redundant cues within television content, though not all of them are readily personalizable. Semantic context is already present in most dramatic content, providing built-in redundancy. Through an ability to control the levels of different objects, or groups of objects, OBA presents an ability to personalize the level of redundant sound cues [13]. More broadly, object-based media broadcast presents the potential to personalize visual objects (e.g., selecting camera angles), or to provide supplementary content to allow viewers to become more familiar with the voices or content of the program.

This work focuses on addressing the sensory needs of those with hearing loss; literature suggests some consideration must be made of cognitive needs when adding additional elements to the content. It has been argued that the effort required for those with hearing loss to filter out background sound and "clean up" the speech means that there

is reduced attention for the higher level cognitive processing required to utilize complementary intelligibility cues [95]. This concept is echoed in a 2000 study by Moreno and Mayer which addressed the effect of additional audio elements on knowledge transference in multimedia learning [96]. This work showed that for instructional messages, additional audio elements can overload the listeners' working memory [96]. A more recent study has shown that for infants, whose cognitive processes are not yet fully developed, music interferes with transfer learning from television content [97]. A 2010 study by Aramaki et al. showed that categorization of ambiguous non-speech sounds takes longer than for typical sounds [98]. These works suggest that the amount of redundant information and how it is presented requires personalization to ensure that an optimal balance between improved intelligibility and cognitive needs is maintained.

3.4 Preliminary Discussion

From the literature three dimensions of personalization for OBA have been identified: speech to noise ratio, spatial separation, and redundancy. Considering the definition of intelligibility at the beginning of this section, increasing speech to noise ratio constitutes an improvement in signal-based intelligibility, making a greater portion of the original speech signal available for the listener. Spatial separation also achieves this to some extent as well as offering complementary spatial cues for the listener. All are evidenced to provide useful functions in overcoming the sensory barriers experienced by viewers. However, for some viewers additional redundant information could have the effect of overloading working memory and be detrimental to intelligibility. This interaction between sensory and cognitive media access needs has to be considered in any accessible personalization implementation.

4 OBJECT-BASED PERSONALIZATION

OBA formats can facilitate much improved broadcast access for hard of hearing people. Recent formats, such as Dolby Atmos, MPEG-H, and DTS:X have the facility to broadcast individual sound elements as independent audio objects, complete with metadata to instruct the receiver as to how the objects should be rendered [99]. OBA can also facilitate personalization of audio presentation based on individual viewer preferences, sensory or environmental needs. This section systematically outlines projects where this functionality has been used to provide end-user personalization of audio, first generally and then with specific applications for hearing impaired listeners.

4.1 Object-Based Audio Personalization

Although initial focus of OBA was on facilitating immersive periphonic (with height) sound, its potential for personalization has become viewed as increasingly important. Personalization possibilities proposed have included alternate sports commentaries [100, 101], home and away sport crowd ambience choices [102], balance between foreground and background sound [103–105], and alternate language provision [106, 107, 105].

Prior to the first broadcast of OBA formats, work using the Web Audio API, had been carried out to explore how personalization could be employed. The BBC and Fraunhofer carried out "Netmix" in 2011, an experiment using a live broadcast of the Wimbledon Tennis Championships that allowed end-users to select between seven options for relative level between commentary and court ambience [108, 109]. Two distinct patterns were apparent in listeners' preferences; slightly less commentary, to enhance the feeling of "*being there*," and considerably more commentary, to improve intelligibility. Similar trials with Swedish Radio content have also been conducted [108].

Other work utilizing the Web Audio API has been completed by the BBC. Personalized dynamic range control, based on the end-user's preferences, needs, and listening environment was proposed and informally evaluated [110]. A demonstration of OBA and the audio definition model using the Web Audio API proposes the ability for the user to mute and unmute individual objects as well as select binaural or stereo rendering in the browser [111].

Further live broadcast experiments by the BBC with football content using the Web Audio API allowed viewers to customize which team's end the crowd noise came from [102]. This work utilized three audio streams: on pitch sounds, commentary, and crowd noise streams, and users could select between predetermined mixes. An interesting result from this study was that two-thirds of participants chose to increase crowd noise relative to commentary. Other experiments using football broadcast material were undertaken as part of the FascinatE project [112, 113]. The FascinatE project featured user-manipulation of the visual point of view of a 180 degree 8K panoramic video, accompanied by corresponding transformations of the audio scene to match. The final demonstration of the project also featured separate user-controls to personalize levels of on-pitch, crowd, and commentary sounds during a game [78], although formal user-testing of this was not carried out.

Automatic selection of a background level to optimize intelligibility has been investigated by Tang et al. [114]. The system utilizes an objective intelligibility metric to analyze the intelligibility of the speech. If the intelligibility falls below a predefined threshold, the system exploits the separation between speech and other sounds in OBA, to adjust the overall speech to background ratio. The threshold at which adjustment occurs could be personalized.

Personalization to improve the listening experience in non-ideal environments and mobile devices has been explored by Walton et al. [103, 104] by allowing users to adjust the foreground/background balance. Background sounds were defined as diffuse and ambient sounds, foreground sounds as dialog and prominent sound effects. Environmental noise had a significant effect on the mix preferences of participants. Again, the results highlighted two distinct clusters of behavior: one tended towards raising the foreground effects to make them audible above the environmental noise and the other increased the background noise to try and mask the environmental noise.

As part of the S3A project, Demonte et al. also investigated personalization potential for mobile and small screen devices [115]. This work explored the effect on intelligibility of binaural auralization of noise, speech or both, as well as the effect of visual information. The study utilized the GRID audio visual corpus [116] with head tracked binaural reproduction to perceptually locate the speech, noise or both, at an external screen. Results indicated a 9.2% increase in intelligibility in the condition with speech externalized to the screen and masker noise reproduced in stereo in the headphones. This effect has been attributed to binaural release from masking [117] and audio-visual coherency when the speech appears to be coming from the speaker on the screen.

More recent work on the S3A project has used OBA to integrate ad hoc arrays of personal and mobile devices into immersive audio reproduction. Media Device Orchestration (MDO) [118] utilizes additional metadata and allows individual objects, or object categories, to be sent to connected mobile devices. MDO enables mobile devices to augment a sound scene to improve immersion and also has potential for specific objects, e.g., narration, to be sent to a specific individual's device.

The Orpheus Project has explored the user experience of object-based media [105]. The project first evaluated perceived usefulness of features before and after use, including the capacity for the viewer to alter: listening perspective, language, audio rendering format including binaural, and the foreground/background balance. Results showed the greatest increase in perceived usefulness was for the foreground/background balance. A user experience study with a large cohort was also evaluated under different listening scenarios: airplane cabin and living room. These tests evaluated different features including different audio reproduction, dynamic range control, and an additional transcript. In the airplane cabin scenario, 83% of participants indicated a preference for binaural reproduction compared to stereo or mono. Over 70% of participants from all age groups also indicated a positive effect in the use of dynamic range control. Improved intelligibility was rated the second best feature overall by participants. Only half the participants found the additional transcript useful. However, participant feedback such as *"I can't hear so well anymore. The transcript would make listening to the radio easier for me"* by a 60-year-old participant, demonstrates that this feature is useful for a subset of listeners.

4.2 Audio Personalization for Accessibility

Early work to address the needs of hard of hearing listeners was conducted by Fraunhofer and termed Spatial Audio Object Coding for Dialogue Enhancement (SAOC-DE) [99, 119]. SAOC-DE was designed to complement existing 5.1 and stereo broadcast systems and transmitted un-mixing metadata that could separate audio objects from the audio mix [99]. In intelligibility tests using the Oldernburg Sentence test [48] and applause style background noise, it was demonstrated that SAOC-DE improved sentence recognition accuracy from 34% to 81%. Related dialog enhancement work for archival content has continued utilizing blind source separation techniques to extract audio objects [120]. This uses the MPEG-H format to facilitate end-user personalization of the speech level [121, 107].

DTS has also presented a dialog-based personalization solution [122]. The proposed algorithm specifies dialog control and enhancement. Alongside this is a protection mechanism to ensure appropriate levels of dialog compared to other program content that is maintained through sections where levels change substantially. The algorithm makes use of object loudness metadata.

The BBC has developed companion screen technology that can deliver an audio description track from a synchronized device [123]. This allows viewers with sight loss to have some control over the audio description and forms the basis of delivery for other types of personalized OBA to an individual listener. Such an approach requires headphones that isolates the listeners from the communal experience of watching television. A solution to providing individualized audio while maintaining the communal experience has been proposed by Simon Galvez et al. [124]. This proposal utilizes highly directional beam-forming, implemented in a consumer-style soundbar, to deliver personalized audio to only the listener requiring it while providing a standard audio mix to additional listeners.

It has been highlighted that the potentially large number of audio objects in a television program, and the fact that OBA allows hypothetical control over all objects, means that a better understanding of the role of these objects and how they can be grouped is required [6]. Work by Woodcock et al. [125] has investigated how people cognitively categorize different parts of broadcast audio for a range of program material. They found that at least seven categories were perceived: continuous and transient background sound, clear speech, non-diegetic music and effects, sounds indicating the presence of people, sounds indicating actions and movement, and prominent attention-grabbing transient sounds. This categorization scheme has been utilized in a reduced form in subsequent work where users were given control of four sound categories: dialog, music, foreground effects, and background effects [14]. This project, a collaboration between DTS and the University of Salford, presented hard of hearing participants with an interface allowing them to adjust the volume of each category. In general, the participants reduced the non-speech categories relative to speech, although the speech itself was left close to its initial level by almost all participants. However, there was substantial inter-personal variation in the levels set for other categories, but lower intra-personal variation across genres. Interestingly, around a third of the participants set levels of the *foreground effects* significantly higher than *music* and *background effects*. In questionnaire responses and discussion these participants stated that the *foreground effects* helped them to understand the media content.

Work investigating object-based personalization for hard of hearing listeners is limited, though interest in the area is increasing [126, 127].

5 DISCUSSION

OBA presents a clear opportunity to improve broadcast experiences for all listeners, not only those with hearing loss. Research focus has primarily been on speech to noise ratio rather than spatial separation and redundancy dimensions. This section will discuss each dimension in turn, highlighting effective strategies as well as unexploited potential. The challenges and opportunities of audio personalization for consumers and the industry are outlined.

5.1 Speech to Noise Ratio

Speech to noise ratio is the most commonly implemented dimension [108, 104, 122, 114, 78, 109, 105]. It is clearly a desirable and effective personalization parameter, given the balances chosen by even normal hearing listeners. It is also powerful in that it can be leveraged to improve intelligibility [119], combat adverse listening environments [103–105], and increase immersion [108].

SOAC-DE, and it's transition into MPEG-H strategies [121], highlight an important question; how to offer personalization for legacy and other pre-mixed content. Such hybrid strategies that make use of the increasing efficacy of source separation algorithms [128, 129] and the personalization capabilities of OBA represent a transitional solution between linear and object-based broadcasting. The integration of producer constraints into metadata allows a balance to be struck between production, broadcaster, and end-users' requirements [122].

Personalized dynamic range control and compression [110] has potential to intelligently interact with hearing aids to prevent problems deriving from multi-band compression being applied multiple times [11]. This is an unexplored area that has significant potential as the ability for devices to communicate directly to hearing aids improves. While the exploitation of hearing aid technology to provide personalization is a complementary area of study, it is outside the scope of this review.

This dimension has the advantage of being conceptually simple for the end-user. It is easily implemented as either a single control [104, 103] or a selection between multiple predefined mixes [102]. Automatic adjustment shows potential to lower this barrier further, particularly if the intelligibility threshold can be set for individual listeners, listening scenarios or preferences [114]. While significant work has explored personalization of speech to noise ratio, automatic adjustment represents a unexploited area.

While conceptually simple for the end-user, this simplicity relies on utilizing the target speech vs. masker (everything else) paradigm [99, 108] or ad hoc definitions of foreground and background sound [103, 104]. However, the distinction between useful and masking sounds is more complex. This is evidenced by the effect of redundant non-speech information on intelligibility [90, 91, 89] and the personal preferences reported by Shirley et al. [14]. While Woodcock et al. demonstrate some generality in people's categorization of broadcast sounds [125], the usefulness or masking potential is not generalizable.

5.2 Spatial Separation

This dimension has received comparatively limited research attention and may be in part due to the majority of terrestrial broadcast utilizing stereo. However, the increase of video on demand consumption and headphone listening may provide the necessary impetus for change. A number of object-based technologies in development have the potential to offer this type of personalization through binaural rendering [111, 105, 115] or sending alternate mixes to secondary devices [118, 123]. The use of binaural rendering has the additional advantage of providing useful location cues to listeners who may also have some sight loss [130]. Transaural soundbar technology has potential to provide personalized speech to noise ratios and also increased spatial separation for individual listeners [124]. As highlighted by Demonte et al., the effect of binaural auralization on the intelligibility of speech is not well known and further research is required [115].

Beyond the technological challenges, spatial separation presents a parameter which is conceptually difficult to personalize. In order to go beyond simple provision of binaural or transaural reproduction, an exploration into adaptation of audio object location and its impact on audiovisual congruency is needed to enable accessible user control. Such a control could take the form of a "spread out" button, personalizing levels of spatial separation, and could be beneficial for hearing and visually-impaired listeners.

5.3 Redundancy

Multi-modal redundancy cues, in the form of subtitles, have reached near ubiquity as an access service in some regions. However, personalization of other redundancy cues, particularly single mode cues, has seen less exploitation. Multi-modal informational redundancy provided by transcripts [105] have seen positive feedback from some users but are yet to be explored in specific accessibility applications. A project carried out by the BBC, called *Story Explorer*, developed tools to create additional online content for media. The content allowed users to explore information relevant to a program such as to story-lines, key events, and characters [60]. This approach, coupled with tools for synchronized second screen content [131], offers technology to deliver supporting information. This could be leveraged to provide additional audio that would allow users to familiarize themselves with voices of characters. What remains to be evaluated is the provision of content structure and accessibility for specific sensory needs.

A cursory investigation by Shirley et al., indicated how single mode redundancy, such as relevant non-speech sounds, can be personalized and exploited [14]. However, as highlighted in Sec. 5.1, the challenge is then how to categorize these sounds in terms of their relative usefulness and masking potential while making personalization easily accessible to the end-user. To address this, Ward and Shirley have proposed a hierarchy of *narrative importance* for audio objects, allowing objects to be categorized by their influence on the story and complementary intelligibility potential rather than by type of sound [132, 133]. In this

approach, each audio object in a sound scene is tagged with *narrative importance* metadata so that the end user can personalize the level of different categories while appropriately balancing levels of those vital to narrative comprehension. This, in effect, gives a sliding scale which end-users can adjust to reduce the complexity of content and therefore of the cognitive load required to process it. Tools to define this narrative importance metadata and to adjust the complexity of the audio reproduction have been implemented and demonstrated by the S3A project [132, 133].

5.4 Commercial Challenges and Opportunities

Implementation of any OBA personalization may have substantial implications for production workflows and for production costs [134]. The BBC's Responsive Radio experiment, which created an object-based variable length radio documentary, took considerable resources to create [135], though efforts are underway to develop tools that move the process into a scalable workflow [59]. In addition to much-needed tool development, interactions between different dimensions of personalization, and between personalization and other access services, are currently undetermined. User-assessment is required to ensure any implemented strategies are effective.

Despite potential shifts in workflow and tools, object-based production opportunities for broadcasters are considerable. The demand for personalized accessible broadcast is only likely to increase with an aging demographic [4]. Positive participant response to trials of object-based accessibility personalization from early work [108] to the most recent [105] and current trends toward Hollywood-style sound design, which many listeners find problematic [24], will drive demand for new access services.

At the time of writing, OBA technology has been rolled out in several territories. Initial broadcasts have been limited in their exploitation of the personalization potential of OBA [136, 107]. MPEG-H is being broadcast in South Korea as the sole audio codec for the country's terrestrial UHDTV broadcasting system [136] and includes facility for audio description (also known as video description services), and for dialog, to be broadcast as audio objects and to be available for personalization [137]. Dolby Atmos broadcast commenced in the UK in January 2017 and, although initially focused on immersive audio for live sport, there are plans to introduce accessible audio features [138]. DTS has demonstrated potential OBA personalization prototypes using object-categorization for hearing impaired people [14]. This demonstrates a drive from audiences and companies to implement object-based broadcast technology for a personalized listening experience.

6 CONCLUSIONS

It is evident no single solution will address all problems faced by individuals with hearing impairments in accessing broadcast audio. However several approaches covered in this review show promise. For legacy and other channel-based media, advances in speech enhancement techniques such as those discussed by Torcoli et al. [139 can be facilitated by OBA. Speech separation algorithms such as these could also be informed by intelligibility metering and adaptation as described in [114] to automate for optimum intelligibility. However the pre-mixed nature of legacy content means that for real improvements to accessible personalization, we need to look to the future.

6.1 Future Directions

This review paper has formalized the otherwise disparate development of OBA personalization strategies for hard of hearing listeners into three dimensions of personalization—speech to noise ratio, spatial separation, and redundancy. From this we can determine which areas are best pursued based on likelihood of adoption and greatest benefit to the target population. A greater research focus on speech to noise ratio is likely due to its simplicity, both in implementation and conceptualization by the end-user. This dimension is likely to have the most widespread implementation and appeal, given that several territories already broadcast in OBA formats capable of personalizing this without substantial further technological development. Limited explorations have been made into personalizing the spatial separation of objects, due to the potential complexity of personalizing this dimension. Delivery of an alternate binaural mix, which inherently contains spatial separation benefits, seems the most likely implementation of this dimension to be broadly adopted, given its existing use to provide more immersive mixes. Both for research and implementation, use of redundant information presents the most interesting unanswered questions. In particular, how redundant information can be leveraged to improve accessibility without excessive cognitive load. It is the opinion of the authors that strategies which blend personalizable speech to background ratio and variable amounts of redundant information are likely to yield the greatest accessibility improvement for hard of hearing listeners. Such a strategy has been developed by the authors as part of the S3A project [133, 132]. This work utilized the concept of narrative importance to group audio objects and thus facilitate easy level adjustment of the different objects through use of a sliding *complexity* scale (further described in [140]).

Implementing personalized solutions is not without challenges, both technological and in designing personalization interfaces which are themselves, accessible and user-friendly. Regardless of dimension, the greatest impact for the majority of audiences will be achieved by strategies that do not rely on specific hardware, such as second screen devices or hearing aids, but have the capacity to interface with additional devices if they are available. How much control over the mix should be ceded to the audience to balance creative integrity and accessibility is one of many unanswered questions. However for broadcasters, the benefits of catering to more consumers with associated increased audience shares are compelling.

It is evident from this work that for broadcast to be truly accessible, it needs to not only exploit OBA technology and accommodate the idiosyncrasies of different types of

hearing loss, but meet the audience's desire for greater agency in an ever expanding media landscape.

7 ACKNOWLEDGMENTS

Lauren Ward is funded by the General Sir John Monash Foundation. The authors would like to thank the contributions of M. Armstrong, P. Demonte, and Dr O. Bones.

8 REFERENCES

[1] Action on Hearing Loss, "Hearing Matters Report" (2015), https://www.actiononhearingloss.org.uk/how-we-help/information-and-resources/publications/research-reports/hearing-matters-report/.

[2] Y. Agrawal, E. A. Platz, and J. K. Niparko, "Prevalence of Hearing Loss and Differences by Demographic Characteristics among US Adults: Data from the National Health and Nutrition Examination Survey, 1999–2004," *Archives of Internal Medicine*, vol. 168, no. 14, pp. 1522–1530 (2008), https://doi.org/10.1001/archinte.168.14.1522.

[3] B. Shield, "Evaluation of the Social and Economic Costs of Hearing Impairment," (2006 Oct.), https://www.hear-it.org/sites/default/files/hear-it%20documents/Hear%20It%20Report%20October%202006.pdf.

[4] Office for National Statistics, "National Population Projections: 2014-Based Statistical Bulletin" (2015 Oct.), https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationprojections/bulletins/nationalpopulationprojections/2015-10-29#older-people.

[5] T. N. Roth, D. Hanebuth, and R. Probst, "Prevalence of Age-Related Hearing Loss in Europe: A Review," *Eur. Arch. Otorhinolaryngol.*, vol. 268, no. 8, pp. 1101–1107 (2011), https://doi.org/10.1007/s00405-011-1597-8.

[6] M. Armstrong, "BBC White Paper WHP 324: From Clean Audio to Object Based Broadcasting" (2016 Oct.), http://www.bbc.co.uk/rd/publications/whitepaper324.

[7] The Nielsen Company (US), "The Total Audience Report Q1, 2017" (2017).

[8] Broadcasters Audience Research Board, "Trends in Television Viewing 2017" (2018 Feb.), https://www.barb.co.uk/download/?file=/wp-content/uploads/2018/03/BARB-Trends-in-Television-Viewing-2017.pdf.

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

[10] D. Cohen, "Sound Matters, BBC College of Production" (2011 Mar.), http://www.bbc.co.uk/academy/production/article/art20130702112136134.

[11] O. Strelcyk and G. Singh, "TV Listening and Hearing Aids," *PloS one*, vol. 13, no. 6 (2018), https://doi.org/10.1371/journal.pone.0200083.

[12] M. Armstrong and M. Crabb, "Exploring Ways of Meeting a Wider Range of Access Needs through Object-Based Media—Workshop," presented at the *Conference on Accessibility in Film, Television and Interactive Media* (2017 Oct.).

[13] L. A. Ward, "Accessible Broadcast Audio Personalisation for Hard of Hearing Listeners," *Adjunct Publication of the 2017 ACM International Conference on Interactive Experiences for TV and Online Video*, pp. 105–108 (2017 Jun.), https://doi.org/10.1145/3084289.3084293.

[14] B. G. Shirley, M. Meadows, F. Malak, J. S. Woodcock, and A. Tidball, "Personalized Object-Based Audio for Hearing Impaired TV Viewers," *J. Audio Eng. Soc.*, vol. 65, pp. 293–303 (2017 Apr.), https://doi.org/10.17743/jaes.2017.0005.

[15] K. Ellis, "Television's Transition to the Internet: Disability Accessibility and Broadband-Based TV in Australia," *Media Intl. Australia*, vol. 153, no. 1, pp. 53–63 (2014), https://doi.org/10.1177/1329878X1415300107.

[16] Cisco, VNI, "Cisco Visual Networking Index: Forecast and Methodology 2016–2021 (2017)" (2017).

[17] ITU - G3ICT, "Digital Inclusion: Making Television Accessible Report" (2011 Nov.), http://staging.itu.int/en/ITU-D/Digital-Inclusion/Persons-with-Disabilities/Documents/Making_TV_Accessible-English.pdf.

[18] "Royal Charter for the Continuance of the British Broadcasting Corporation" (2016 Dec.), http://downloads.bbc.co.uk/bbctrust/assets/files/pdf/about/how_we_govern/2016/charter.pdf.

[19] Australian Government, "Australian Broadcasting Corporation Act 1983: Compilation no. 28" (2018 18 Mar.), https://www.legislation.gov.au/Details/C2018C00079.

[20] V. Nationen, "Convention on the Rights of Persons with Disabilities (CRPD)," *Resolution*, vol. 61, p. 106 (2016), https://www.un.org/development/desa/disabilities/convention-on-the-rights-of-persons-with-disabilities/convention-on-the-rights-of-persons-with-disabilities-2.html.

[21] D. Monzani, G. Galeazzi, E. Genovese, A. Marrara, and A. Martini, "Psychological Profile and Social Behaviour of Working Adults with Mild or Moderate Hearing Loss," *Acta Otorhinolaryngol. Ital.*, vol. 28, no. 2, p. 61 (2008).

[22] H. Fullerton, "BBC Drama SS-GB Criticised for 'Mumbling' and Bad Sound Quality in First Episode" (2017 26 Feb.), http://www.radiotimes.com/news/2017-02-26/bbc-drama-ss-gb-criticised-for-mumbling-and-bad-sound-quality-in-first-episode.

[23] J. Plunkett, "Heard this Before? BBC Chief Speaks Out over Happy Valley Mumbling" (2016 8 Apr.), https://www.theguardian.com/media/2016/apr/08/bbc-happy-valley-mumbling-jamaica-inn-sarah-lancashire.

[24] Hanard, "Television Broadcasts: Audibility" (2017 4 Apr.), https://hansard.parliament.uk/lords/2017-04-04/debates/F84C55A0-3D8B-41F7-A19C-CC216F8C7B0B/TelevisionBroadcastsAudibility.

[25] C. D. Mathers, "A Study of Sound Balances for the Hard of Hearing," NASA STI/Recon Technical Report N, vol. 91 (1991).

[26] Voice of the Listener and Viewer, "VLV's Audibility of Speech on Television Project Will Make a Real Difference" (2011 Jun.), http://www.vlv.org.uk/documents/06.11PressreleasefromVLV-AudibilityProject-0800hrs1532011_002.pdf.

[27] P. Mapp, "Intelligibility of Cinema & TV Sound Dialogue," presented at the *141st Convention of the Audio*

Engineering Society (2016 Sep.), convention paper 9632, http://www.aes.org/e-lib/browse.cfm?elib=18436.

[28] Access Economics, "Listen Hear! The Economic Impact and Cost of Hearing Loss in Australia," *Report for The Cooperative Research Centre for Cochlear Implant and Hearing Aid Innovation and Victorian Deaf Society* (2006).

[29] W. H. Organization, et al., "Deafness and Hearing loss," *Fact sheet*, vol. 300 (2015).

[30] G. A. Gates and J. H. Mills, "Presbycusis," *The Lancet*, vol. 366, no. 9491, pp. 1111–1120 (2005), https://doi.org/10.1016/S0140-6736(05)67423-5.

[31] K. T. Palmer, M. J. Griffin, H. E. Syddall, A. Davis, B. Pannett, and D. Coggon, "Occupational Exposure to Noise and the Attributable Burden of Hearing Difficulties in Great Britain," *J. Occup. Environ. Med.*, vol. 59, no. 9, pp. 634–639 (2002), https://doi.org/10.1136/oem.59.9.634.

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

[33] M. Maassen, W. Babisch, K. D. Bachmann, H. Ising, G. Lehnert, P. Plath, P. Plinkert, E. Rebentisch, G. Schuschke, and M. Spreng, et al., "Ear Damage Caused by Leisure Noise," *Noise and Health*, vol. 4, no. 13, p. 1 (2001).

[34] R. J. H. Smith, A. E. Shearer, M. S. Hildebrand, and G. Van Camp, "Deafness and Hereditary Hearing Loss Overview" (1993).

[35] P. M. Rabinowitz, "Noise-Induced Hearing Loss," *Amer. Family Physician*, vol. 61, no. 9, pp. 2759–2760 (2000).

[36] British Society of Audiology, "Recommended Procedure: Pure-Tone Air-Conduction and Bone-Conduction Tthreshold Audiometry with and without Masking" (2011).

[37] A. Bronkhorst and R. Plomp, "Effect of Multiple Speech-Like Maskers on Binaural Speech Recognition in Normal and Impaired Hearing," *J. Acoust. Soc. Amer.*, vol. 92, no. 6, pp. 3132–3139 (1992), https://doi.org/10.1121/1.404209.

[38] M. K. Pichora-Fuller, B. A. Schneider, and M. Daneman, "How Young and Old Adults Listen to and Remember Speech in Noise," *J. Acoust. Soc. Amer.*, vol. 97, no. 1, pp. 593–608 (1995), https://doi.org/10.1121/1.412282.

[39] Australia. Parliament. Senate. Community Affairs References Committee. and Siewert, Rachel. *Hear us : inquiry into hearing health in Australia / The Senate Community Affairs References Committee* The Senate Community Affairs References Committee Canberra 2010

[40] H. Aazh, D. Prasher, K. Nanchahal, and B. C. J. Moore, "Hearing-Aid Use and its Determinants in the UK National Health Service: A Cross-Sectional Study at the Royal Surrey County Hospital," *Int. J. Audiol.*, vol. 54, no. 3, pp. 152–161 (2015), https://doi.org/10.3109/14992027.2014.967367.

[41] A. J. Vermiglio, S. D. Soli, D. J. Freed, and L. M. Fisher, "The Relationship between High-Frequency Pure-Tone Hearing Loss, Hearing in Noise Test (HINT) Thresholds, and the Articulation Index," *J. Amer. Acad. of Audiol.*, vol. 23, no. 10, pp. 779–788 (2012), https://doi.org/10.3766/jaaa.23.10.4.

[42] M. Huckvale and G. Hilkhuysen, "On the Predictability of the Intelligibility of Speech to Hearing Impaired Listeners," presented at the *1st International Workshop on Challenges in Hearing Assistive Technology* (2017 Aug.).

[43] R. Plomp, "Auditory Handicap of Hearing Impairment and the Limited Benefit of Hearing Aids," *J. Acoust. Soc. Amer.*, vol. 63, no. 2, pp. 533–549 (1978), https://doi.org/10.1121/1.381753.

[44] D. Ruggles and B. Shinn-Cunningham, "Spatial Selective Auditory Attention in the Presence of Reverberant Energy: Individual Differences in Normal-Hearing Listeners," *J. Assoc. Res. Otolaryngol.*, vol. 12, no. 3, pp. 395–405 (2011), https://doi.org/10.1007/s10162-010-0254-z.

[45] C. Füllgrabe, B. C. J. Moore, and M. A. Stone, "Age-Group Differences in Speech Identification Despite Matched Audiometrically Normal Hearing: Contributions from Auditory Temporal Processing and Cognition," *Front. Ageing Neurosci.*, vol. 6 (2015), https://doi.org/10.3389/fnagi.2014.00347.

[46] M. Nilsson, S. D. Soli, and J. A. Sullivan, "Development of the Hearing in Noise Test for the Measurement of Speech Reception Thresholds in Quiet and in Noise," *J. Acoust. Soc. Amer.*, vol. 95, no. 2, pp. 1085–1099 (1994), https://doi.org/10.1121/1.408469.

[47] R. A. McArdle, R. H. Wilson, and C. A. Burks, "Speech Recognition in Multitalker Babble Using Digits, Words, and Sentences," *J. Amer. Acad. Audiol.*, vol. 16, no. 9, pp. 726–739 (2005), https://doi.org/10.3766/jaaa.16.9.9.

[48] K. Wagener, V. Kühnel, and B. Kollmeier, "Development and Evaluation of a German Sentence Test I: Design of the Oldenburg Sentence Test," *Zeitschrift Fur Audiologie*, vol. 38, pp. 4–15 (1999).

[49] K. Hopkins and B. C. J. Moore, "The Contribution of Temporal Fine Structure to the Intelligibility of Speech in Steady and Modulated Noise," *J. Acoust. Soc. Amer.*, vol. 125, no. 1, pp. 442–446 (2009), https://doi.org/10.1121/1.3037233.

[50] E. Villchur, "Simulation of the Effect of Recruitment on Loudness Relationships in Speech," *J. Acoust. Soc. Amer.*, vol. 56, no. 5, pp. 1601–1611 (1974), https://doi.org/10.1121/1.1903484.

[51] R. Badri, J. H. Siegel, and B. A. Wright, "Auditory Filter Shapes and High-Frequency Hearing in Adults Who Have Impaired Speech in Noise Performance Despite Clinically Normal Audiograms," *J. Acoust. Soc. Amer.*, vol. 129, no. 2, pp. 852–863 (2011), https://doi.org/10.1121/1.3523476.

[52] C. J. Plack, D. Barker, and G. Prendergast, "Perceptual Consequences of 'Hidden' Hearing Loss," *Trends in Hearing*, vol. 18, p. 2331216514550621 (2014), https://doi.org/10.1177/2331216514550621.

[53] R. Schaette and D. McAlpine, "Tinnitus with a Normal Audiogram: Physiological Evidence for Hidden Hearing Loss and Computational Model," *J. Neurosci.*, vol. 31,

no. 38, pp. 13452–13457 (2011), https://doi.org/10.1523/JNEUROSCI.2156-11.2011.

[54] J. R. Dubno, D. D. Dirks, and D. E. Morgan, "Effects of Age and Mild Hearing Loss on Speech Recognition in Noise," *J. Acoust. Soc. Amer.*, vol. 76, no. 1, pp. 87–96 (1984), https://doi.org/10.1121/1.391011.

[55] A. Wingfield, P. A. Tun, and S. L. McCoy, "Hearing Loss in Older Adulthood: What it Is and How it Interacts with Cognitive Performance," *Curr. Dir. Psychol. Sci.*, vol. 14, no. 3, pp. 144–148 (2005), https://doi.org/10.1111/j.0963-7214.2005.00356.x.

[56] F. Marmel, D. Linley, R. Carlyon, H. Gockel, K. Hopkins, and C. Plack, "Subcortical Neural Synchrony and Absolute Thresholds Predict Frequency Discrimination Independently," *J. Assoc. Res. Otolaryngol.*, vol. 14, no. 5, pp. 757–766 (2013), https://doi.org/10.1007/s10162-013-0402-3.

[57] C. G. Clinard and K. L. Tremblay, "Aging Degrades the Neural Encoding of Simple and Complex Sounds in the Human Brainstem," *J. Amer. Acad. Audiol.*, vol. 24, no. 7, pp. 590–599 (2013), https://doi.org/10.3766/jaaa.24.7.7.

[58] L. Tolstoy, *Anna Karenina*, vol. 2 (1966).

[59] J. Cox, M. Brooks, I. Forrester, and M. Armstrong, "Moving Object-Based Media Production from One-Off Examples to Scalable Workflows," *SMPTE Motion Imaging J.*, vol. 127, no. 4, pp. 32–37 (2018), https://doi.org/10.5594/JMI.2018.2806499.

[60] M. Evans, T. Ferne, Z. Watson, F. Melchior, M. Brooks, P. Stenton, and I. Forrester, "Creating Object-Based Experiences in the Real World," presented at the *International Broadcast Convention* (2016), https://doi.org/10.1049/ibc.2016.0034.

[61] L. Fontan, J. Tardieu, P. Gaillard, V. Woisard, and R. Ruiz, "Relationship between Speech Intelligibility and Speech Comprehension in Babble Noise," *J. Speech, Lang. Hear. Res.*, vol. 58, no. 3, pp. 977–986 (2015), https://doi.org/10.1044/2015_JSLHR-H-13-0335.

[62] IEC60268-16," Standard, International Electrotechnical Commission (2011).

[63] N. Miller, "Measuring Up to Speech Intelligibility," *Int. J. Lang. Comm. Disorders*, vol. 48, no. 6, pp. 601–612 (2013), https://doi.org/10.1111/1460-6984.12061.

[64] D. N. Kalikow, K. N. Stevens, and L. L. Elliott, "Development of a Test of Speech Intelligibility in Noise Using Sentence Materials with Controlled Word Predictability," *J. Acoust. Soc. Amer.*, vol. 61, no. 5, pp. 1337–1351 (1977), https://doi.org/10.1121/1.381436.

[65] A. A. Zekveld, M. Rudner, I. S. Johnsrude, D. J. Heslenfeld, and J. Rönnberg, "Behavioral and fMRI Evidence that Cognitive Ability Modulates the Effect of Semantic Context on Speech Intelligibility," *Brain Lang.*, vol. 122, no. 2, pp. 103–113 (2012), https://doi.org/10.1016/j.bandl.2012.05.006.

[66] B. Lindblom, "On the Communication Process: Speaker-Listener Interaction and the Development of Speech," *Augment. Alt. Comm.*, vol. 6, no. 4, pp. 220–230 (1990), https://doi.org/10.1080/07434619012331275504.

[67] J. Bizley and Y. Cohen, "The What, Where and How of Auditory-Object Perception," *Nature Rev. Neurosci.*, vol. 14, no. 10, pp. 693–707 (2013), https://doi.org/10.1038/nrn3565.

[68] S. Coren, "Most Comfortable Listening Level as a Function of Age," *Ergonomics*, vol. 37, no. 7, pp. 1269–1274 (1994), https://doi.org/10.1080/00140139408964905.

[69] J. K. Willcox, "Better TV Sound for Those With Hearing Loss https://www.consumerreports.org/lcd-led-oled-tvs/better-tv-sound-for-those-with-hearing-loss/" (2018 12 Mar).

[70] A. Carmichael, "Evaluating Digital 'On-Line' Background Noise Suppression: Clarifying Television Dialogue for Older, Hard-of-Hearing Viewers," *J. Neuropsychol. Rehab.*, vol. 14, no. 1-2, pp. 241–249 (2004), https://doi.org/10.1080/09602010343000192.

[71] TVC, "D4.4 — Pilot-B Evaluations and Recommendations," Tech. Rep. (2016), http://pagines.uab.cat/hbb4all/sites/pagines.uab.cat.hbb4all/files/d4.4-tvc_pilot-b-evaluations-and-recommendations_v1.00.pdf.

[72] M. Armstrong, "Audio Processing and Speech Intelligibility: A Literature Review," *BBC Research & Development Whitepaper* (2011).

[73] I. J. Hirsh, "The Relation between Localization and Intelligibility," *J. Acoust. Soc. Amer.*, vol. 22, no. 2, pp. 196–200 (1950), https://doi.org/10.1121/1.1906588.

[74] T. L. Arbogast, C. R. Mason, and G.Jr Kidd, "The Effect of Spatial Separation on Informational Masking of Speech in Normal-Hearing and Hearing-Impaired Listeners," *J. Acoust. Soc. Amer.*, vol. 117, no. 4, pp. 2169–2180 (2005), https://doi.org/10.1121/1.1861598.

[75] 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.).

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

[77] T. Holman, *Sound for Film and Television* (Focal Press, 2012), https://doi.org/10.4324/9780240814322.

[78] B. Shirley, *Improving Television Sound for People with Hearing Impairments*, Ph.D. thesis, University of Salford (2013).

[79] O. Strelcyk, G. Singh, L. Standaert, L. Rakita, P. Derlath, and S. Launer, "TV/Media Listening and Hearing Aids," *Poster presented at: International Hearing Aid Conference* (2016).

[80] M. Armstrong, A. Brown, M. Crabb, C. J. Hughes, R. Jones, and J. Sandford, "Understanding the Diverse Needs of Subtitle Users in a Rapidly Evolving Media Landscape," presented at the *International Broadcast Convention Conference* (2015), https://doi.org/10.1049/ibc.2015.0032.

[81] M. Pluymaekers, M. Ernestus, and R. Baayen, "Articulatory Planning Is Continuous and Sensitive to Informational Redundancy," *Phonetica*, vol. 62, no. 2-4, pp. 146–159 (2005), https://doi.org/10.1159/000090095.

[82] R. Bilger, Speech Recognition Test Development, in E. Elkins, ed., *Speech Recognition by the Hearing Impaired*, vol. 14, pp. 2–15 (1984).

[83] R. H. Wilson, R. McArdle, K. L. Watts, and S. L. Smith, "The Revised Speech Perception in Noise Test (R-SPIN) in a Multiple Signal-to-Noise Ratio Paradigm," *J. Amer. Acad. Audiol.*, vol. 23, no. 8, pp. 590–605 (2012), https://doi.org/10.3766/jaaa.23.7.9.

[84] D. J. Schum and L. J. Matthews, "SPIN Test Performance of Elderly Hearing-Impaired Listeners," *J. Amer. Acad. Audiol.*, vol. 3, no. 5, pp. 303–307 (1992).

[85] L. E. Humes, B. U. Watson, L. A. Christensen, C. G. Cokely, D. C. Halling, and L. Lee, "Factors Associated with Individual Differences in Clinical Measures of Speech Recognition among the Elderly," *J. Speech, Lang. Hear. Res.*, vol. 37, no. 2, pp. 465–474 (1994), https://doi.org/10.1044/jshr.3702.465.

[86] L. Ward, B. Shirley, Y. Tang, and W. Davies, et al., "The Effect of Situation-Specific Non-Speech Acoustic Cues on the Intelligibility of Speech in Noise," presented at *Interspeech 2017* (2017), https://doi.org/10.21437/Interspeech.2017-500.

[87] P. Souza, N. Gehani, R. Wright, and D. McCloy, "The Advantage of Knowing the Talker," *J. Amer. Acad. Audiol.*, vol. 24, no. 8, pp. 689–700 (2013), https://doi.org/10.3766/jaaa.24.8.6.

[88] R. L. Freyman, U. Balakrishnan, and K. S. Helfer, "Effect of Number of Masking Talkers and Auditory Priming on Informational Masking in Speech Recognition," *J. Acoust. Soc. Amer.*, vol. 115, no. 5, pp. 2246–2256 (2004), https://doi.org/10.1121/1.1689343.

[89] N. Hodoshima, "Effects of Urgent Speech and Preceding Sounds on Speech Intelligibility in Noisy and Reverberant Environments," *Interspeech 2016*, pp. 1696–1699 (2016), https://doi.org/10.21437/Interspeech.2016-1618.

[90] L. Ward, B. Shirley, Y. Tang, and W. Davies, "The Effect of Situation-Specific Acoustic Cues on Speech Intelligibility in Noise," *Interspeech 2017*, pp. 2958–2962 (2017 Aug.), https://doi.org/10.21437/Interspeech.2017-500.

[91] L. Ward and B. Shirley, "Television Dialogue; Balancing Audibility, Attention and Accessibility," presented at the *Conference on Accessibility in Film, Television and Interactive Media* (2017 Oct.).

[92] M. Aguert, V. Laval, L. Le Bigot, and J. Bernicot, "Understanding Expressive Speech Acts: The Role of Prosody and Situational Context in French-Speaking 5- to 9-Year-Olds," *J. Speech, Lang. Hear. Res.*, vol. 53, no. 6, pp. 1629–1641 (2010), https://doi.org/10.1044/1092-4388(2010/08-0078).

[93] A. A. Zekveld, M. Rudner, I. S. Johnsrude, J. M. Festen, J. H. M. Van Beek, and J. Rönnberg, "The Influence of Semantically Related and Unrelated Text Cues on the Intelligibility of Sentences in Noise," *Ear Hear.*, vol. 32, no. 6, pp. 16–25 (2011), https://doi.org/10.1097/AUD.0b013e318228036a.

[94] B. Spehar, S. Goebel, and N. Tye-Murray, "Effects of Context Type on Lipreading and Listening Performance and Implications for Sentence Processing," *J. Speech, Language, and Hearing Res.*, vol. 58, no. 3, pp. 1093–1102 (2015), https://doi.org/10.1044/2015_JSLHR-H-14-0360.

[95] 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.

[96] R. Moreno and R. E. Mayer, "A Coherence Effect in Multimedia Learning: The Case for Minimizing Irrelevant Sounds in the Design of Multimedia Instructional Messages." *J. Ed. Psychol.*, vol. 92, no. 1, p. 117 (2000), https://doi.org/10.1037//0022-0663.92.1.117.

[97] R. Barr, L. Shuck, K. Salerno, E. Atkinson, and D. L. Linebarger, "Music Interferes with Learning from Television during Infancy," *Infant Child Dev.*, vol. 19, no. 3, pp. 313–331 (2010), https://doi.org/10.1002/icd.666.

[98] M. Aramaki, C. Marie, R. Kronland-Martinet, S. Ystad, and M. Besson, "Sound Categorization and Conceptual Priming for Nonlinguistic and Linguistic Sounds," *J. Cogn. Neurosci.*, vol. 22, no. 11, pp. 2555–2569 (2010), https://doi.org/10.1162/jocn.2009.21398.

[99] J. Paulus, J. Herre, A. Murtaza, L. Terentiv, H. Fuchs, S. Disch, and F. Ridderbusch, "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.

[100] S. Meltzer, M. Neuendorf, D. Sen, and P. Jax, "MPEG-H 3D Audio—The Next Generation Audio System," presented at the *International Broadcast Convention* (2014), https://doi.org/10.1049/ib.2014.0011.

[101] S. Mehta and T. Ziegler, "Personalized and Immersive Broadcast Audio," presented at the *International Broadcast Convention* (2014), https://doi.org/10.1049/ib.2014.0010.

[102] M. Mann, A. W. P. Churnside, A. Bonney, and F. Melchior, "Object-Based Audio Applied to Football Broadcasts," *ACM International Workshop on Immersive Media Experiences*, pp. 13–16 (2013), https://doi.org/10.1145/2512142.2512152.

[103] T. Walton, M. Evans, D. Kirk, and F. Melchior, "Does Environmental Noise Influence Preference of Background-Foreground Audio Balance?" presented at the *141st Convention of the Audio Engineering Society* (2016 Sep.), convention paper 9637.

[104] T. Walton, M. Evans, D. Kirk, and F. Melchior, "Exploring Object-Based Content Adaptation for Mobile Audio," Personal Ubiquitous Comput., pp. 1–14 (2018). http://doi.org/10.1007/s00779-018-1125-6

[105] D5.6: Report on Audio subjective tests and user tests" (2018 6 July), https://orpheus-audio.eu/wp-content/uploads/2018/07/orpheus-d5.6_report-on-audio-subjective-and-user-tests_v1.3.pdf.

[106] R. Bleidt, A. Borsum, H. Fuchs, and S. M. Weiss, "Object-Based Audio: Opportunities for Improved Listening Experience and Increased Listener Involvement," *SMPTE Motion Imaging J.*, vol. 124, no. 5, pp. 1–13 (2015 July), https://doi.org/10.5594/j18579.

[107] R. Brun, "Successful Demonstration of Interactive Audio Streaming Using MPEG-H Audio at Norwegian Broadcaster NRK" (2018 2 July), /www.audioblog.iis.fraunhofer.com/mpegh-nrk/.

[108] H. Fuchs, S. Tuff, and C. Bustad, "Dialogue Enhancement—Technology and Experiments," *EBU Technical Review*, vol. 2 (2012).

[109] H. Fuchs and D. Oetting, "Advanced Clean Audio Solution: Dialogue Enhancement," *SMPTE Motion Imaging J.*, vol. 123, no. 5, pp. 23–27 (2014), https://doi.org/10.5594/j18429.

[110] A. Mason and M. Paradis, "Adaptive, Personalised 'in Browser' Audio Compression," presented at the *1st Web Audio Conference* (2015).

[111] C. Pike, P. Taylour, and F. Melchior, "Delivering Object-Based 3D Audio Using the Web Audio API and the Audio Definition Model," presented at the *1st Web Audio Conference* (2015).

[112] FP7/2007-2013, Grant agreement no. 248138" (2010), www.fascinate-project.eu.

[113] 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.), https://doi.org/10.17743/jaes.2015.0017.

[114] Y. Tang, B. M. Fazenda, and T. J. Cox, "Automatic Speech-to-Background Ratio Selection to Maintain Speech Intelligibility in Broadcasts Using an Objective Intelligibility Metric," *Appl. Sci.*, vol. 8, no. 1, p. 59 (2018), https://doi.org/10.3390/app8010059.

[115] P. Demonte, Y. Tang, R. J. Hughes, T. Cox, B. Fazenda, and B. Shirley, "Speech-to-Screen: Spatial Separation of Dialogue from Noise Towards Improved Speech Intelligibility for the Small Screen," presented at the *144th Convention of the Audio Engineering Society* (2018 May), convention paper 10011.

[116] M. Cooke, J. Barker, S. Cunningham, and X. Shao, "An Audio-Visual Corpus for Speech Perception and Automatic Speech Recognition," *J. Acoust. Soc. Amer.*, vol. 120, no. 5, pp. 2421–2424 (2006), https://doi.org/10.1121/1.2229005.

[117] H. Levitt and L. Rabiner, "Binaural Release from Masking for Speech and Gain in Intelligibility," *J. Acoust. Soc. Amer.*, vol. 42, no. 3, pp. 601–608 (1967), https://doi.org/10.1121/1.1910629.

[118] J. Francombe, J. Woodcock, R. J. Hughes, R. Mason, A. Franck, C. Pike, T. Brookes, W. J. Davies, P. J. Jackson, and T. J. Cox, et al., "Qualitative Evaluation of Media Device Orchestration for Immersive Spatial Audio Reproduction," *J. Audio Eng Soc.*, vol. 66, pp. 414–429 (2018 Jun.), https://doi.org/10.17743/jaes.2018.0027.

[119] H. Fuchs and D. Oetting, "Advanced Clean Audio Solution: Dialogue Enhancement," presented at the *IBC2013 Conference Proceedings* (2013 Sep.), https://doi.org/10.1049/ibc.2013.0002.

[120] A. Craciun, C. Uhle, and T. Bäckström, "An Evaluation of Stereo Speech Enhancement Methods for Different Audio-Visual Scenarios," *23rd European Signal Processing Conference (EUSIPCO)*, pp. 2048–2052 (2015), https://doi.org/10.1109/EUSIPCO.2015.7362744.

[121] M. Torcoli, J. Herre, H. Fuchs, J. Paulus, and C. Uhle, "The Adjustment/Satisfaction Test (A/ST) for the Evaluation of Personalization in Broadcast Services and Its Application to Dialogue Enhancement," IEEE Transact. Broadcast. (2018), https://doi.org/10.1109/TBC.2018.2832458.

[122] 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.

[123] V. Vinayagamoorthy, R. Ramdhany, and M. Hammond, "Enabling Frame-Accurate Synchronised Companion Screen Experiences," *ACM International Conference on Interactive Experiences for TV and Online Video*, pp. 83–92 (2016), https://doi.org/10.1145/2932206.2932214.

[124] Gálvez M. F. Simón, S. J. Elliott, and J. Cheer, "Time Domain Optimization of Filters Used in a Loudspeaker Array for Personal Audio," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 23, no. 11, pp. 1869–1878 (2015), https://doi.org/10.1109/TASLP.2015.2456428.

[125] J. S. Woodcock, W. J. Davies, and T. J. Cox, F. Melchior, "Categorization of Broadcast Audio Objects in Complex Auditory Scenes," *J. Audio Eng Soc.*, vol. 64, pp. 380–394 (2016 Jun.), https://doi.org/10.17743/jaes.2016.0007.

[126] T. R. Agus and C. Corrigan, "Adapting Audio Mixes for Hearing Impairments," presented at the *3rd Workshop on Intelligent Music Production* (2018 Sep.).

[127] I. Fraile, J. A. Nuñez J. A. N. Mario Montagud, and S. Fernández, "ImAc: Enabling Immersive, Accessible and Personalized Media Experiences," *2018 ACM International Conference on Interactive Experiences for TV and Online Video*, pp. 245–250 (2018 Jun.).

[128] A. Ephrat, I. Mosseri, O. Lang, T. Dekel, K. Wilson, A. Hassidim, W. T. Freeman, and M. Rubinstein, "Looking to Listen at the Cocktail Party: A Speaker-Independent Audio-Visual Model for Speech Separation," arXiv preprint arXiv:1804.03619 (2018), https://doi.org/10.1145/3197517.3201357.

[129] W03—Audio Repurposing Using Source Separation" (2018 May), http://www.aes.org/events/144/workshops/?ID=5917.

[130] M. Lopez and G. Kearney, "Enhancing Audio Description: Sound Design, Spatialisation and Accessibility in Film and Television," presented at the *Proc. of Institute of Acoustics 32nd Reproduced Sound Conf.* (2016 Nov.).

[131] D2.1 System Architecture" (2017 14 July), https://2immerse.eu/wp-content/uploads/2018/01/d2.1_r2-system_architecture-clean.pdf.

[132] Accessible Audio Research: UK Council on Deafness Presentation" (2017 Nov.), http://usir.salford.ac.uk/47781/1/12.%20Lauren%20Ward%20%26%20Ben%20Shirley.mp4.

[133] L. Ward and B. Shirley, "Intelligibility vs Comprehension: Understanding Quality of Accessible Next-Generation Audio Broadcast" (2018), URL http://usir.salford.ac.uk/id/eprint/47780.

[134] P. Poers, "Challenging Changes for Live NGA Immersive Audio Production," presented at the *144th Convention of the Audio Engineering Society* (2018 May), eBrief 418.

[135] M. Armstrong, M. Brooks, A. Churnside, M. Evans, F. Melchior, and M. Shotton, "Object-Based Broadcasting-Curation, Responsiveness and User

Experience," presented at the *International Broadcasting Convention* (2014), https://doi.org/10.1049/ib.2014.0038.

[136] Standard KO-07.0127R1: TT—Transmission and Reception for Terrestrial UHDTV Broadcasting Service, Revision 1" (2016 Dec.).

[137] R. L. Bleidt, D. Sen, A. Niedermeier, B. Czelhan, S. Füg, S. Disch, J. Herre, J. Hilpert, M. Neuendorf, H. Fuchs, J. Issing, A. Murtaza, A. Kuntz, M. Kratschmer, F. Küch, R. Füg, B. Schubert, S. Dick, G. Fuchs, F. Schuh, E. Burdiel, N. Peters, and M. Y. Kim, "Development of the MPEG-H TV Audio System for ATSC 3.0," *IEEE Transact. Broadcast.*, vol. 63, no. 1, pp. 202–236 (2017 Mar.), https://doi.org/10.1109/TBC.2017.2661258.

[138] J. Riedmiller, S. Mehta, N. Tsingos, and P. Boon, "Immersive and Personalized Audio: A Practical System for Enabling Interchange, Distribution, and Delivery of Next-Generation Audio Experiences," *SMPTE Motion Imaging J.*, vol. 124, no. 5, pp. 1–23 (2015), https://doi.org/10.5594/j18578.

[139] M. Torcoli, J. Herre, J. Paulus, C. Uhle, H. Fuchs, and O. Hellmuth, "The Adjustment/Satisfaction Test (A/ST) for the Subjective Evaluation of Dialogue Enhancement," presented at the *143rd Convention of the Audio Engineering Society* (2017 Oct.), convention paper 9842.

[140] L. Ward, B. Shirley, and J. Francombe, "Accessible Object-Based Audio Using Hierarchical Narrative Importance Metadata," presented at the *145th Convention of the Audio Engineering Society* (2018 Oct.), eBrief 478.

THE AUTHORS

Lauren A. Ward

Ben G. Shirley

Lauren Ward is a Postgraduate Researcher in broadcast accessibility at the Acoustics Research Centre, University of Salford, UK. She has a B.Eng. with First Class Honours (2014) and a B.Phil. (2015) from the University of Tasmania, Australia. Lauren currently works on object-based broadcasting and accessible audio technology for hard of hearing listeners, for which she won the EPSRC Connected Nation Pioneer Award for *making digital technology work better for people*. She is a General Sir John Monash Scholar and 2018 recipient of the IET's Leslie H. Paddle Postgraduate Scholarship. Previously Lauren has worked with the CSIRO developing speech recognition technology for the assessment of speech disorders in children. Her research interests also include STEM education, ASR, and machine learning.

•

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 and contributed to clean audio methods documented in ETSI, EBU, and other digital broadcast standards as part of the UK Clean Audio Forum. Dr. Shirley's current research activity is in future spatial audio systems, audio content analysis, automated audio mixing for broadcast, and application of object-based audio for accessibility.