Improving broadcast accessibility for hard of hearing individuals

using object-based audio personalisation and narrative importance

Lauren Alison Ward

Supervisor: Dr. Ben Shirley
Prof. Bill Davies

Acoustics Research Centre
School of Computing, Science and Engineering

Doctorate of Philosophy in Acoustics and Audio Engineering

University of Salford May 2020

"We're all born mad. Some remain so."

– Samuel Beckett, Waiting for Godot

This work is dedicated to the memory of Gordon *'Tere'* Manns,

who never let an idea pass him without interrogating it.

Abstract

Technological advances in broadcasting can be the impetus for advances in accessibility services. For the 11 million individuals in the United Kingdom with some degree of hearing loss, the advent of object-based broadcasting and it's personalisation features has the potential to facilitate a transition towards more accessible broadcast audio. Part I of this work conducts a systematic review of previous object-based accessibility research, identifying the personalisation of redundant non-speech objects as a potentially high impact yet unexplored area of research. Guided by these findings, and the results of a survey of end-user needs, the specific research questions of this work are then developed as:

1. What is the relationship between redundant non-speech audio objects and broadcast speech intelligibility, for normal and hard of hearing listeners?
2. Can a system be designed which allows end-users to control the balance between audio objects for dramatic content which is simple to use and preserves comprehension?

Part II of this work shows that the presence of redundant non-speech sounds improve speech recognition in noise in normal hearing listeners, even when the sound partially masks the speech. Subsequent investigations show that this effect exists within hard of hearing cohorts also, and the benefit yielded by non-speech sounds can be predicted by the severity of hearing loss in an individual's better hearing ear.

Part III work translates these novel findings into practical broadcast accessibility technology, through the development of a new conceptual framework called: 'Narrative Importance'. Based on this framework, production tools and an end-user interface are developed and deployed in a large scale public trial. The results of this trial demonstrate that this new approach to accessible audio can deliver content which is more enjoyable with reduced energetic masking of speech, whilst still maintaining the creative integrity and comprehension of the content.

Acknowledgements

Research is far from an individual achievement and there are so many friendly giants who have let me stand on their shoulders to get here; those who have contributed to this research in ways big and small, weird and wonderful.

First of all, to my fellow inhabitants of the fish bowl (and those who managed escaped). In particular thank you Will, for always being willing to dig through R code with me, and Philippa for being my speech intelligibility buddy and the greatest experiment day organiser the world has ever seen.

To the acoustics research centre lab staff, past and present: for last minute experiment participation, lending audiometers and long chats about stats over oreos.

To the production staff who have let me pick their brains: Martyn Whitley, Eloise Whitmore, Howard Bargroff and Martyn Harries. Without you I would not know the narrative importance of the door in the 'Rover's Return'.

To the members of the S3A Project, thank you for adopting me and providing the foundations upon which much this research has been built.

To BBC R&D Northlab, for being the most brilliant workplace and for providing an endless supply of ~~happiness~~ cake on the kitchen table. In particular, my immense gratitude goes to the the audio team. Working with you has turned the crazy, un-achievable goals I set for myself at the beginning of this PhD into a reality.

Particular thanks go to Jon Francombe for creating the first version of the NI control, colloquially known as "the knob". Thank you for making the ideas we had into something real and tangible. To Matthew Paradis, for creating the second NI control and the producer experiment interface – for your tireless hours spent on wrangling the WebAudio API, I am truly grateful. To Rick Hughes for the first NI metadata plug-in and to Manos Chourdakis,

for his work on the second. To my dear sister-in-law Katherine Tucker, for many hours spent transcribing IPA and Catherine Robinson for re-recording all the R-SPIN sentences.

Ginormous thanks go to Laura Russon, Rhys Davies, Gabriella Leon and the Casualty Post Production team. Your enthusiasm for trying new things made the A&E Audio trial not only possible, but a joy to be part of.

My gratitude goes to those who put their money behind this project: the Institute of Acoustics, EPSRC, ISCA, ACM, the IET, BBC R&D and most of all the General Sir John Monash Foundation. Without the Sir John Monash Foundation and the belief and guidance of Peter Binks and Judith Landsberg, none of this would have been possible.

To those who have edited and applied copious amounts red-pen to this work I, and the subsequent readers of this thesis, thank you: Chris Baume, Chris Pike, Hannah Clawson, Cammi Motley, Matteo Torcoli Bill Davies and Will Bailey. And most of all to those who have critiqued this work cover to cover: Pamela Ward, Kate Tucker, Mike Armstrong and Ben Shirley.

And finally those special few

My family, who instilled in me the curiosity and the wanderlust that led me to a PhD on the other side of the world and cheered me on throughout it.

Mike Armstrong for all the *looong* discussions and kitchen table editing sessions. You attract interesting people into your orbit and I am proud to count myself among them.

Ben Shirley – more colleague and friend, than supervisor. It has been a joy to work with you, busk presentations with you and drink fine tequila with you. Thank you for rolling the dice on a random Australian, thank you for introducing me to everyone worth knowing and thank you for being an excellent human being.

And to Matthew Tucker – my champion and the sane, steady presence in my frenetic brainstorm. We are better as us. From the bottom of my heart, **thank you**.

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this thesis are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation contains fewer than 100,000 words including appendices.

This thesis is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the following section, Collaborative work. Sections of collaborative work are also noted throughout the thesis text as footnotes at the beginning of the relevant sections of work. Sections of work in this thesis which had been published at the time of writing are noted in the List of Publications.

Lauren Alison Ward
May 2020

Collaborative work

- Chapter 3
  - In Section 3.4.4 coding of free text responses was undertaken as part of the directed content analysis of the survey results. This was undertaken by two additional researchers to improve the reliability of the coding: Dr. Ben Shirley and Dr. Alex Wilson. All analysis of the codings was undertaken by the author.
- Chapter 7
  - The voice recordings for the R^2SPIN in Section 7.4.1 were made by Catherine Robinson. All other work in developing the R^2SPIN was completed by the author.
  - In Section 7.4.2, the phonetic transcription of the original and re-recorded R-SPIN sentences was undertaken by Katherine Tucker. The summaries of these transcriptions and all subsequent usage of the transcriptions were performed by the author.
- Chapter 8
  - The data collection in Section 8.2.2 and 8.2.3 was collected by the author and a team of additional researchers: Dr. Lara Harris, Philippa Demonte, Zuza Podwinska, Dr. William Bailey, Georgia Shirley and Dr. Ben Shirley. All analysis of these data was completed by the author.
- Chapter 9
  - The initial NI control prototype, described in Section 9.3 was developed collaboratively by Dr. Ben Shirley, Dr. Jon Francombe and the author, as part of the S3A: Future Spatial Audio for the Home project. Dr. Jon Francombe coded the function and interface for this prototype and performed the NI metadata assignment to the audio objects in 'The Turning Forest' and the 'Protest' scene used in Section 9.4.
  - In Section 9.4, the focus groups were facilitated by Dr. Ben Shirley, with notes taken by Philippa Demonte. Coding of the data was undertaken by an additional researcher to improve the reliability of the coding: Dr. Ben Shirley. All analysis of the codings was undertaken by the author.
- Chapter 10
  - The initial version of the NI metadata assignment plug-in, described in Section 10.2.1 was developed as part of the S3A: Future Spatial Audio for the Home project. It was coded by Dr. Rick Hughes, with support from Dr. Jon Fran-

combe and Dr. James Woodcock. All work resulting from the use of this plug-in described in Section 10.2 was completed by the author.

  - The interface for the online metadata assignment task in Section 10.3.1 was collaboratively designed by the author and Dr. Matthew Paradis, and implemented in the WebAudio API by Dr. Matthew Paradis. All other data collection and analysis undertaken in Section 10.3 was completed by the author.

- Chapter 11
  - The trial described in Chapter 11 was a collaboration between the author, the University of Salford, BBC R&D and BBC Studios' Casualty programme.
  - The content described in Section 11.2.1 was generated by BBC Studios' 'Casualty' and used with their permission.
  - The NI mix of the episode described in Section 11.3 was completed by Laura Russon, with input from the author, Rhys Davies and Dafydd Llewelyn. The description of the mixing process was completed by the author and verified by Laura Russon. The analysis of the mix was completed by the author.
  - The integration of the NI slider to the Standard Media Player described in Section 11.2.2 was completed by Dr. Matthew Paradis based on the author's guidance and with support from Robin Moore. The scripts which logged interaction data from participants was scripted by Matthew Paradis. All analysis of the interaction data was completed by the author.
  - The pre-roll video explaining the use of the slider was written by Gabriella Leon, with support from the author, and produced by Joe Marshall on the Casualty set.
  - Publicity of the trial was supported by Steve Alderton and Joe Marshall.

Contents

List of Publications

Publications are listed with reference to the chapter reporting the same work. Works reporting preliminary results are noted.

Chapter 2

- **Lauren Ward** and Ben Shirley. Personalization in object-based audio for accessibility: A review of advancements for hearing impaired listeners. *Journal of the Audio Engineering Society*, 67(7/8), 2019. https://doi.org/10.17743/jaes.2019.0021
- Ben Shirley and **Lauren Ward**. Intelligibility vs. comprehension: Understanding quality of accessible next-generation audio broadcast. *Universal Access in the Information Society*, Special Issue on “Quality of Media Accessibility Products and Services", 2019. [Accepted; In press]

Chapter 3

- **Lauren Ward**, Ben Shirley, and Alex Wilson. ‘Audibility and accessibility: Understanding what audiences need from television audio.’ *Journal of the Audio Engineering Society*, 2019. [Under Review]

Chapter 4

- **Lauren Ward**, Ben G Shirley, and William J Davies. ‘Big pictures and little screens: How television sound research can work with, and for, hard of hearing viewers.’ In *33rd Reproduced Sound*. Institute of Acoustics, Nottingham, UK, Nov, 2017.

Chapter 5

- **Lauren Ward**, Ben Shirley, Yan Tang, and William J. Davies. 'The effect of situation specific non-speech acoustic cues on the intelligibility of speech in noise.' In *Interspeech 2017, 18th Annual Conference of the International Speech Communication Association*, pg. 2958-2962, ISCA, Stockholm, Sweden, Aug. 2017. https://doi.org/10.21437/Interspeech.2017-500
- **Preliminary Results:**
  - **Lauren Ward**, Ben Shirley, and William J Davies. 'Turning up the background noise; the effects of salient non-speech audio elements on dialogue intelligibility in complex acoustic scenes.' In *32nd Reproduced Sound*. Institute of Acoustics, Southampton, UK, Nov. 2016.

Chapter 6

- **Lauren Ward** and Ben G Shirley. 'Television dialogue; balancing audibility, attention and accessibility.' In *Conference on Accessibility in Film, Television and Interactive Media*, York, UK, 2017.

Chapter 7

- **Lauren Ward**, Catherine Robinson, Matthew Paradis, Katherine Tucker, and Ben Shirley. 'R^2SPIN: re-recording the revised speech perception in noise test.' In *INTERSPEECH 2019, 20th Annual Conference of the International Speech Communication Association*, pg. 3133-3137, ISCA, Graz, Austria, Sept. 2019. http://dx.doi.org/10.21437/Interspeech.2019-1281

Chapter 8

- **Preliminary Results:**
  - **Lauren Ward**, Ben Shirley, and Bill Davies. 'Development and preliminary results of the University of Salford media Accessibility and hearing Impairment Database (U-SAID).' In *34th Reproduced Sound*. Institute of Acoustics, Bristol, UK, Nov. 2018.

Chapter 9

- **Lauren Ward**, Ben Shirley, and Jon Francombe. 'Accessible object-based audio using hierarchical narrative importance metadata.' In *Audio Engineering Society Convention 145*. Audio Engineering Society, New York, USA, Oct. 2018.
- Ben Shirley, **Lauren Ward**, and Emmanouil Chourdakis. 'Personalization of object-based audio for accessibility using narrative importance.' In *DataTV: 1st International Workshop on Data-Driven Personalisation of Television*, Manchester, UK, June 2019.

Chapter 10

- **Lauren Ward**, Maxine Glancy, Sally Bowman and Mike Armstrong 'The impact of new forms of media on production tools and practices.' In *International Broadcast Convention Conference*. Forthcoming, Sept 2020.

Chapter 11

- **Preliminary Results:**
  - **Lauren Ward**, Matthew Paradis, Ben Shirley, Laura Russon, Robin Moore, and Rhys Davies. 'Casualty Accessible and Enhanced (A&E) Audio: Trialling object-based accessible TV audio.' In *Audio Engineering Society Convention 147*. Audio Engineering Society, New York, USA, Oct. 2019.
- Maxine Glancy, **Lauren Ward** and Mike Armstrong 'Object-Based Media –an overview of the user experience.' In *International Broadcast Convention Conference*. Forthcoming, Sept 2020.

Other Publications

These include publications which give an overview of the research and publications which relate to, but are not part of, this PhD.

- Emmanouil Chordakis, **Lauren Ward**, Matthew Paradis, and Josh Reiss. 'Modelling experts' decisions on assigning narrative importances of objects in a radio drama mix.' In *22nd International Conference on Digital Audio Effects - DAFX*, Birmingham, UK, Sept. 2019.
- Marcos Simon Galvez, Ioseb Laghidze, **Lauren Ward**, Andreas Franck, Ben Shirley, and Filippo Fazi.'Multi-zone personalisation for hard of hearing listeners using object-based audio.' In *34th Reproduced Sound*. Institute of Acoustics, Bristol, UK, Nov. 2018.
- **Lauren Ward**. 'Demonstration of a novel audio customisation system to improve broadcast accessibility for hard of hearing listeners.' In *SPARC 2018 Internationalisation and collaboration: Salford postgraduate annual research conference book of abstracts*. University of Salford, Salford, UK, July. 2018.
- **Lauren Ward**. 'Accessible broadcast audio personalisation for hard of hearing listeners.' In *Adjunct Publication of the 2017 ACM International Conference on Interactive Experiences for TV and Online Video*, ACM. pages 105–108, Hilversum, Netherlands, June, 2017. https://doi.org/10.1145/3084289.3084293
- **Lauren Ward**. 'Snap, crackle, pop: How sound effects help, and hinder, hearing in broadcast audio.' In *SPARC 2017 retrospect & prospects: Salford postgraduate annual research conference book of abstracts*. University of Salford, Salford, UK, July. 2017.

List of Figures

List of Tables

Part I

The Questions

Chapter 1

Motivations

Being hard of hearing isn't a lot of fun but [audio personalisation] puts us back in control. We are in charge of what we hear and I think that's quite empowering.

– Focus group participant, 13/07/2018[1]

One in every six people in the UK is either hard of hearing or D/deaf [2] [3]. Likewise one in six Australians [4] (2006) and Americans [5] (2003-2004) have some degree of hearing loss. The largest cause of this loss is age related; due to ageing populations [6, 3, 7], it is projected that by 2035 the number of individuals with hearing loss in the UK will rise to 15.6 million or one in every five people [3]. Broadcast accessibility services are vital to ensure that this increasing segment of the population can effectively access audiovisual content. With such a large and diverse group of individuals, these services cannot assume that *'one size fits all'*: they must be adaptable and personalisable to the needs of individual consumers.

Not only is the prevalence of hearing loss increasing, those in the demographic most susceptible to hearing loss consume the largest amount of broadcast content. On average, those over 65 years of age in the UK watch more than 5.5 hrs of television daily [8]. In the United States, viewers over 55 years old also consistently watch more hours of television than their younger counterparts [9]. Even accounting for viewing online and on mobile devices which dominates consumption patterns for younger demographics, those over 65 in the UK still watch on average 1 hr more per day of audiovisual content than those aged 16-34 [10].

[1] Quote from hard of hearing focus group participant. The results from this focus group as described in Chapter 9.

[2] The term D/deaf is used throughout this text. This term encompasses two groups of individuals: those who belong to the Deaf community and those who are audiologically deaf but do not identify with Deaf culture [1]. Cultural Deafness *is linked to a social construction of identity, involvement with a Deaf community, a concept of Deaf culture and the use of BSL (British Sign Language).*[2].

Beyond this, access to television is valued by hearing impaired individuals and can provide vital social inclusion. The United Nations Convention on the Rights of Persons with Disabilities Articles 9 and 21 emphasise that access to information, communication and mass media services for those with disabilities is a human right [11]. The importance of accessing content for education, entertainment and national identity is recognised by international standards bodies such as the ITU [12] and echoed in the charters of national broadcasters [13, 14]. The majority (84%) of the hearing-impaired participants in a recent study reported that hearing well when watching TV/video was 'very' or 'extremely' important [15]. Coupled with increased rates of depression and and social isolation amongst adults with even mild to moderate hearing loss [16], it becomes a social imperative to provide the requisite broadcast accessibility services for them.

It is necessary for provision of these services to go beyond terrestrial broadcasting which, due to technological advances, is now but one of many content providers. YouTube alone reports that we watch over 1 billion hours of online videos a day [17]. It is projected that by 2021 audiovisual content will make up 82% of all internet traffic [18]. As we increasingly gain news and information and share our culture through many varieties of audiovisual content, the need for more personal and diverse access services becomes more critical.

These technological changes present challenges but also opportunities. At the time of writing, broadcasting technology is undergoing a fundamental paradigm shift. This shift moves broadcasting from the linear, channel-based model which has been utilised for decades to what is termed *'next generation audio'* [19, 20]. This change presents an opportunity for broadcast accessibility to be drastically improved for hard of hearing listeners [21]. The next generation of audio, rather than being fundamentally structured around the transmission of audio channels, treats individual sounds as independent 'objects'. This allows the objects to be kept separate during broadcast and are reconstructed at the point of service using each object's associated metadata [22–24]. Given this, the end-user's television set can reproduce the objects differently, altering the number of objects and the level at which they are reproduced. This technological advance establishes the groundwork for truly personalisable television audio which can adapt for not only taste but for the specifics of a user's hearing needs.

The last big shift in broadcasting was the switch over from analogue to digital broadcasting during the 2000s. This shift brought with it increased broadcasting capacity and capability to transmit 5.1 channel surround sound whilst serving as an impetus for new accessible audio strategies [25, 26]. The transition to object-based broadcasting is only in the early stages of implementation by the broadcast industry. As this shift occurs, it is vital that the technology's potential be exploited in a way that considers implementation ease

for broadcasters to ensure uptake. Furthermore, it is crucial that developments meet real end-user needs, balanced with maintaining the creative integrity of content. To create tools and services capable of delivering these aims requires technological development which is both user driven and evidence based.

This work thus becomes a cornerstone for change, making significant contributions to knowledge, which are:

- Expanding Armstrong's 'Audibility Problem Space'[21] to differentiate between *audibility* and *accessibility* challenges (Chapter 3).
- Creation of a novel approach to accessible audio, termed 'Narrative Importance', which works for both production staff and end-users (Chapter 9– 11).
- Understanding of the effects of non-speech sounds with narrative importance on speech understanding and how this varies as a function of hearing loss (Chapter 8).

These contributions are achieved through the following body of work. Part I establishes the research questions this work focuses on. These questions are formed based on learning from past opportunities to improve broadcast audio accessibility (Chapter 2) and by conducting a detailed analysis of end-user's current needs (Chapter 3). Part II is then able to build on this to generate novel understanding of broadcast speech perception for normal and hard of hearing individuals. Collaborating closely with hard of hearing individuals, creatives and broadcasters, this new knowledge is then leveraged to develop a novel audio personalisation tool which balances content integrity, viewer agency and ease of implementation (Part III: Chapter 9, 10 and 11). This work has deepened understanding of audience requirements, developed new knowledge about speech perception in broadcast and assembled a novel tool for audio personalisation based on individual user needs. Together this work represents a significant advancement towards meeting our changing audience's accessibility needs in an ever-expanding media landscape.

Chapter 2

Accessible broadcast audio research: A review

2.1 Introduction

This chapter aims to catalogue and evaluate previous work in broadcast accessibility, for both channel-based and next generation audio systems. This will allow for identification of where progress has been made in previous accessible audio research. It will also enable identification of where future research and development should be focused for accessible broadcast audio technology to have the maximum positive impact for end-users. This chapter will also furnish the reader with the requisite knowledge about broadcast accessibility, hearing loss, and speech intelligibility.

This chapter first presents an overview of hearing loss including its prevalence and characterisation as well as the topology and function of the human ear. A brief review of speech intelligibility concepts precedes an outline on the current barriers which individuals with hearing loss face accessing broadcast content. A chronological review of *'clean audio'* research in channel-based broadcasting follows. Object-based audio, the next generation of broadcast audio technology, is then introduced. A rubric derived from previous studies in human hearing and speech intelligibility is then used to categorise the areas where object-based audio personalisation could achieve greatest end-user impact. A systematic review of all object-based personalisation is undertaken with reference to this rubric. The chapter concludes with a discussion of areas of personalisation where the most impact can be achieved.

2.2 Hearing loss

To achieve effective improvements in audio accessibility for the hard of hearing requires an understanding of the types and characteristics of hearing loss, the challenges they present for speech understanding and the number of people affected. The following section provides the reader with a working understanding of this.

2.2.1 Prevalence

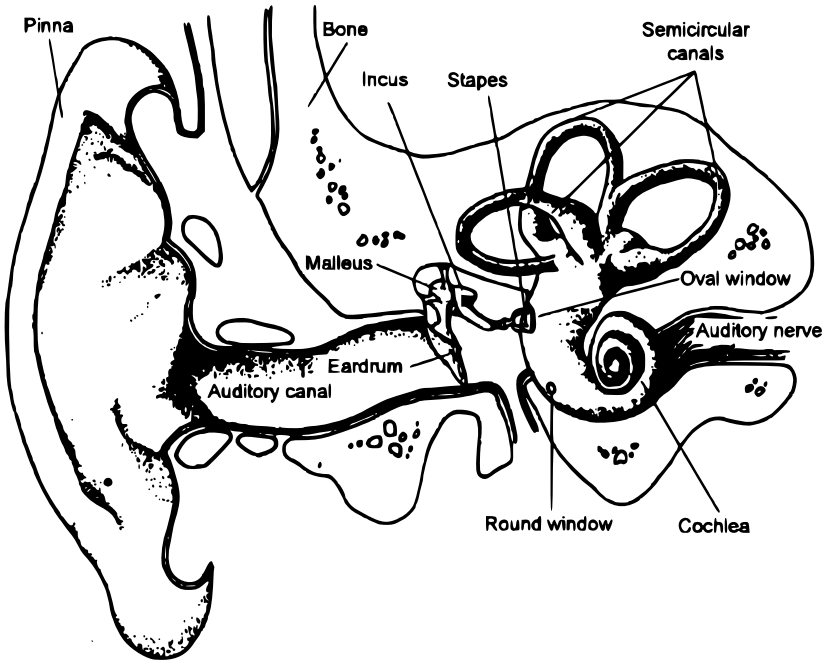
The 'Hearing Matters' report compiled by the hearing loss charity *Action on Hearing Loss* indicates that in 2015 11 million people in the UK were affected by hearing loss [3]. These statistics are mirrored in countries with similar demographics with one in six Australians [4] (2006) and Americans [5] (2003-2004) having some hearing loss. 'Hearing Matters' estimates that in the UK 6.7 million people could benefit from the use of a hearing aid [3]. However generally only a small proportion of these people actually have one fitted (24% in Australia [27]) and many of those who have had a hearing aid fitted do not use them regularly [28, 27]. *Action on Hearing Loss* project that by 2035, the number of individuals with hearing loss in the UK will rise to 15.6 million or one in five people [3]. Given that the single largest cause of hearing loss is age-related loss (Presbycusis [6]), a significant driver of this increase is an ageing population in the UK [3]. Another major cause is noise-induced hearing loss resulting from occupational exposure [29, 30] or from recreational activities such as concerts [31].

2.2.2 Normal function of the ear

The human ear is comprised of three parts: the outer, middle, and inner ear. The parts can be seen in Figure 2.1 (reproduced from [32]). The outer ear is made up of the pinna and meatus (auditory canal). Incoming acoustic signals will resonate in these cavities and this resonance serves to increase the signal's amplitude at the ear drum over the main range of speech frequencies (1.5kHz - 5kHz). Reflections in the pinna, along with reflections from the head and torso, also introduce a complex pattern of peaks and troughs into the incoming signal's spectrum which is leveraged by the auditory system for sound localisation.

The middle ear's main function is to transfer the airborne vibrations received by the outer ear into the fluid borne vibrations of the cochlea. The tympanic membrane (eardrum) transfers the vibrations from the auditory canal through to the ossicles via three small bones called the incus, malleus and stapes in the middle ear. The vibrations are then transferred to an outer membrane of the cochlea called the oval window.

Figure 2.1 Structure of the ear, reproduced from Moore [32].



Two membranes divide the cochlea - the Reissner's membrane, which is a very thin fluid barrier, and the Basilar membrane, which is a multi-cellular structure. Within the cochlea, sitting on the Basilar membrane, is the organ of Corti. This contains the hair cells which have tufts of stereocilia, which look like tiny hairs, at their apexes. Two sets of these hair cells exist: the outer and the inner, which refer to their location relative to the inside and outside of the cochlea's spiral. The main role of the outer hairs cells is to influence and control the mechanics of the cochlea based on external stimuli and higher level control from the auditory system. The inner hair cells serve the primary role of translating mechanical information in the fluid of the cochlea into neural information for higher level auditory processing.

2.2.3 Characterisation of hearing loss

Hearing loss may occur as a result of any part of the peripheral and ascending auditory system not functioning correctly or may be caused by auditory processing problems in the brain, or due to a combination of both. The type of hearing loss an individual may have is characterised by the location of the impairment within the auditory system: *conductive* hearing loss is due to problems within the ear canal, ear drum or middle ear, *sensorineural* hearing loss is due to problems with the inner ear and *mixed* loss is due to both [33]. Conductive loss results in an attenuation of the sounds reaching the cochlea. Sensorineural covers both cochlea

loss, due to damage or incorrect functioning of the cochlea, and neural loss which is due to the incorrect functioning of the auditory nerve. Presbycusis (age-related hearing loss) and noise-induced hearing loss are common causes of sensorineural hearing loss [34].

Table 2.1 Audiometric descriptors based on averages of pure-tone audiometric thresholds at 250, 500, 1000, 2000 and 4000 Hz [35])

Descriptor	**Average Hearing threshold levels (dB HL)**
Mild	20 - 40
Moderate	41 - 70
Severe	71 - 95
Profound	> 95

The audiometric descriptors in Table 2.1 provide simple tools for discussing hearing loss severity. However, it has long been acknowledged that audiometric thresholds alone do not fully account for the variability in individual's ability to understand speech in noise or hear in complex or adverse listening scenarios [36–38]. This difference between audiometric thresholds and speech in noise performance was modelled by Plomp in 1978, who defined two loss classes: *Attenuation*, loss due to the reduction in perceived sound level, as a result of reduced audiometric thresholds and, *Distortion*, which is comparable with a decrease in the effective speech to noise ratio [38]. Other works have termed this *Distortion* class supracochlear loss [37] or suprathreshold loss [39, 40]. The remainder of this work will utilise the term suprathreshold.

2.2.4 Suprathreshold loss

There are numerous contributors to suprathreshold loss which include tinnitus [41, 42], loudness recruitment [43], loss of temporal fine structure [44] or reduction in frequency resolution [45]. To give the readers a working knowledge of the suprathreshold factors affecting hearing ability, they are briefly summarised below. In depth outlines of these factors can be found in the works of Baguley [42] and Moore [32]. [1]

Tinnitus Tinnitus refers to the perception of a tone or multiple tones without the presence of an external acoustic source [42]. It can be persistent, pulsating or intermittent and whilst usually perceived as a tone, can also manifest as noise. It can occur on its own or can be a

[1] There is a lack of agreement between researchers on the terminology and definitions used in referring to different suprathreshold factors [46]. In addition to furnishing the reader with a summary of suprathreshold factors, the following provides a reference for how these terms will be utilised throughout the remainder of this work.

symptom of hearing loss and ear related conditions such as Meniere's disease, head injury [47], depression [48] or as a side effect from ototoxic medication [49] (such as many types of chemotherapy). The tones perceived by those with tinnitus will often vary in loudness based on situation and it is often reported that tinnitus is worsened by stress [42]. There is no cure for tinnitus, though tinnitus induced by stress, anxiety and other psychological conditions can be mitigated by addressing the root cause. Due to the difficulty of establishing a strict definition of tinnitus, results from studies attempting to quanitfy the prevalence of tinnitus vary. However it is estimated that 10-15% of the adult population are affected by tinnitus [50], though it is higher in populations regularly exposed to loud noises e.g. musicians [51]. Tinnitus can affect hearing by interfering with the interpretation of the incoming auditory signal as well as causing listening fatigue, particularly if it is of the persistent variety. It is also commonly co-morbid with hyperacusis (increased sound sensitivity), with around 40% of tinnitus sufferers also experiencing hyperacusis [52].

Loudness perception disorders There are many types of loudness perception disorders, though the most commonly discussed and experienced are hyperacusis [53] and loudness recruitment. Hyperacusis is a reduced tolerance for, or sensitivity to, sounds at a level which does not bother normal hearing listeners [53, 54, 51]. It is surmised that hyperacusis is caused by problems in the peripheral auditory system, diseases of the central nervous system diseases, or hormonal and infectious diseases, though in some cases it has no discernible cause [55].

Loudness recruitment is the effective decrease of tolerable dynamic range meaning individuals experience a reduced acceptable range between their lower loudness threshold (where quiet sounds are imperceptible due to threshold loss) and their upper threshold (where sounds become painfully loud) [56, 43, 57, 32]. This has been demonstrated in an fMRI study by Langers et al. which found that stimulus at the same sound intensity produced higher fMRI activation in the auditory cortex for hearing impaired participants than normal hearing [58], suggesting an altered relationship between sound intensity and perceived loudness. These perception disorders affect the ease at which speech can be understood comfortably, as the level required for speech to be intelligible may cause the listener pain or physical discomfort.

Loss of resolution: temporal and frequency This encompasses the loss of temporal resolution, reduced sensitivity to temporal fine structure and reduced frequency selectivity. The causes of these conditions are not well understood and it is not known whether they can

be attributed to deterioration of central auditory processing or peripheral auditory function, or both [59].

Frequency selectivity, or frequency resolution, refers to the ability of the auditory system to separate and resolve the spectral components of a complex sound [32]. When functioning correctly, this allows an individual to discriminate a target signal from masking signals at different frequencies. This is akin to being presented with a complex tone with multiple harmonics and being able to 'tune in' on a particular harmonic [32]. Whilst there is a minimum distance between harmonics that is required for a normal hearing listener to be able to separate out the tones, this distance increases for those with reduced frequency selectivity. Poor frequency selectivity can lead to confusion between speech sounds which occur in similar frequency bands. This differs from frequency discrimination or pitch perception, which relate to the ability to detect changes in frequency over time e.g. how separated in frequency two consecutive tones have to be for a difference to be perceived. Frequency discrimination ability can also affect speech perception [32].

Temporal resolution refers to the ability to detect changes in a signal over time [32]. Effective temporal resolution allows a listener to resolve speech cues including the temporal envelope of speech, the duration of speech segments and the duration of silent intervals [59]. It has been repeatedly demonstrated that hard of hearing and older listeners will have a poorer ability to detect these temporal cues [60–64], particularly when the speech is presented against noise [65]. The temporal resolution of an individual will often be measured using gap detection tasks [64] or through their sensitivity to amplitude modulation [63]. This differs from temporal integration, which is the ability to combine auditory information collected over time to better detect or discriminate a stimuli [32].

Sensitivity to temporal fine structure (TFS) refers to the ability to perceive the rapid oscillations close to the centre frequency of a complex sound, such as speech. TFS can be considered as the carrier of envelope information in an incoming signal. It has been shown to play a significant role in pitch perception [66] as well as being hypothesised to play a role in speech perception in competing noise and in stream segregation [44]. It has also been shown that hearing impaired listeners and eldery normal hearing listeners have a reduced ability to utilise TFS information [67].

Other factors Within clinical practice, any other factors that are not identified by standard measurements of hearing but impacts upon an individual's hearing ability are referred to as hidden hearing loss [68]. In this manner, the majority of suprathreshold factors outlined here would be contributors to hidden hearing loss. In auditory research however, the term is often used more specifically to refer to hearing loss which is hypothesised to be caused by cochlea

neuropathy (the loss of the high-threshold auditory nerve fibres) [69, 70]. Traditionally it has been assumed that after noise exposure, the inner hair cells are the first to be damaged but recent studies suggest that it is the synapses between hair cells and cochlear nerve terminals which degenerates first [71]. This degeneration affects speech perception but its effects are not measurable by an audiogram. Numerous studies have demonstrated this to be the case in mammals, however its presence in humans has remained controversial due to the difficulty in measuring the synaptic function in live humans [72]. Recent studies evaluating the cochlea synapses in postmortem humans who had aged normally add to the evidence suggesting that cochlea neuropathy is a cause of 'hidden' hearing loss [73, 72].

2.3 Speech intelligibility

Most accessible audio research has been predicated on improving the intelligibility of broadcast speech. This section briefly outlines the definition and types of intelligibility.

2.3.1 Definition of speech intelligibility

Speech intelligibility is often defined as the proportion of words which are correctly heard, strictly differentiating it from comprehension [74]. Utilising this definition, speech intelligibility refers to only the perception of the words and is a sensory issue only. Intelligibility can also be defined as the proportion of words understood [75], which incorporates elements of comprehension and quality thus making it both a sensory and cognitive issue. This work will utilise the latter, broader definition as both perception of the speech as well as understanding of it is required for broadcast content to be considered accessible.

The literature distinguishes between two forms of speech intelligibility: *signal-dependent* (using 'bottom-up processing'), where the ability to retrieve the message is based solely on the speech signal, and *complementary*, which utilises other non-speech cues from the speech signal, such as syntax and semantics, as well as multi-modal cues such as facial expressions [76]. These complementary cues are also referred to as 'top-down information' [77] and they have been shown to play a greater role in speech perception when hearing is challenged, either by impairment or masking from competing sources [78, 79]. It is theorised that improvement in intelligibility in the presence of complementary cues is due to the manner in which the brain composes perceptual auditory objects: using expectations and schemata of what it believes the object will be, based on experience and knowledge of language structure, in order to predict parts of the object for which no input signal is currently available [80].

2.3.2 Complementary intelligibility

A significant body of research exists demonstrating the release from masking that the inclusion of complementary intelligibility cues can produce. These cues include, but are not limited to, spatial separation [81], contextual cues [82], priming the listener with the speaker's voice or speech content [83, 84], and multi-modal cues [85, 86]. Complementary intelligibility cues are particularly relevant to accessible broadcast research as television content uses a rich array of these cues: visual context, context through sound effects, dialogue based context as well as access services such as audio description and subtitling (termed captioning or closed captioning in the United States). Complementary intelligibility is described in greater detail with reference to the personalisation dimensions in Sections 2.6.2 and 2.6.2 respectively.

2.4 Current services and challenges

It is essential that the development of any accessible audio technology addresses real problems faced by end-users and does not replicate existing services. This section first outlines the current levels of access services provision in terrestrial, video on-demand and online services followed by the challenges still faced by hard of hearing individuals attempting to engage with broadcast content. A short exploration of the barriers to access experienced in German broadcasting and how they compare to the English language broadcast landscape is also made.

2.4.1 Current access services

Access services are additional provisions or versions of content provided by broadcasters to enable individuals with specific needs, such as sight or hearing loss, to better engage with their content. There are three existing recognised access services in the UK: subtitles, signing and audio description.

Access service provision varies considerably across Europe and although some standardisation processes have taken place, each territory is responsible for mandating levels and types of services [87]. Optional regulations exist at the EU level to encourage distributors to carry any and all available access services (preventing these potentially being stripped from the broadcast to reduce bandwidth), but they require adoption and enforcement by individual countries.

In the UK, access services such as signing and subtitles are mandated by Ofcom across a proportion of programming with the amount of programming that must be accessible

varying according to audience share. For example the BBC, as a public service broadcaster with a large audience share is mandated to, and provides, subtitling across almost 100% of its broadcast output and signing for 5% [88]. France mandates 100% subtitling and sign language for at least 3 news programmes per day [89], Spain mandates subtitling for 45% of private and 55% of public channels and sign language for 1 (private channels) or 3 (public channels) hours per week. A useful summary of current practice can be found in [90] and up to date access service mandates and level of provision in the UK can be found in [88].

At the time of writing, the UK media regulator Ofcom only has the power to enforce subtitling quotas for terrestrial broadcasters. As a result of on-demand and streaming providers' access services being unregulated, a 2015 study by Action on Hearing Loss found that 87% of respondents had attempted to watch a show on-demand only to find it lacked subtitling [91]. The passage of the 'Digital Economy Act 2017' [92] paves the way for Ofcom to enforce subtitling quotes on on-demand and streaming platforms. Ofcom has recently completed a public consultation, with the result being a recommended target of 40% of content subtitled within two years and 80% within four years. Whilst not mandated for online only services, a significant portion of Youtube's content is captioned (one billion captioned videos in 2017 [93]). However, as automatic approaches are used to caption videos which are optionally edited and refined by creators, the quality of these captions vary.

2.4.2 Current barriers to access

For many hearing impaired viewers, particularly those with mild to moderate loss, subtitles and sign language are not an optimal solution to improve the intelligibility of speech on TV. Whilst loudness levels of programming are now strictly defined by standards [94] and recommendations [95], the level of the dialogue as compared to other elements is not. Some broadcasters do make reference to dialogue levels in delivery specifications [96–99] however they are often poorly defined and vary considerably between broadcasters and between programme genres.

A large-scale study in 2008, by the Royal National Institute for the Deaf (the forerunner to Action on Hearing Loss), reported that 87% of hard of hearing viewers struggled to understand speech on television [100]. Similar difficulty was revealed in a cross-sectional survey of a single evening's viewing carried out by BBC in 2010 [101]. This found that 60% of viewers had difficulty hearing what was said in the broadcasts at some point during the evening, highlighting the fact that intelligibility issues are not only experienced by those with hearing loss. It also identified four key factors that make speech hard to understand: clarity of speech, unfamiliar or strong accents, background noise, and background music [101, 102]. When combined, the effect of these factors are compounded [103].

Figure 2.2 'Audibility Problem Space', which describes the points in the broadcast chain where reductions in audibility can occur and the sources of these reductions, reproduced from BBC White Paper 324 [21]

Audibility Problem Space				
Production /Direction	**Capture**	**Post Production**	**Broadcast**	**Home**
Writing style – complexity of narrative Direction and camerawork Choice of shots – showing actor's face at vital moments in the narrative Voices, accents, dialects & clarity of delivery Choice of location - control of background noise	Choice of microphone Microphone positioning Skill of sound recordist Manual level control vs agc/limiter Priority given to sound over video – e.g. microphone allowed in shot Retakes for sound Voiceover booth matching to field recordings - ADR Use of digital compression and codec choice (mobile phone audio!)	Voice processing, equalisation, noise reduction /gating, level control/limiting Added music, level, type, purpose, style Added sound effects, level, type purpose Added voiceover Video edit, choice of shots Loudness control Dynamic range Stereo or 5.1 mix Inclusion of open subtitles	Output processing or other dynamic range control - not currently used on BBC TV Audio encoding quality and any cascaded coding Quality of closed subtitles	Television sound quality – receiver, amplifier and speakers Use of mono, stereo or 5.1 loudspeakers Room Acoustics and the positions of the television and the viewer in the room Background noise Viewer's hearing - level of interest in the programme - knowledge of the programme's subject matter - expectations of the programme - language skills - willingness to use subtitles

In a more recent study of hard of hearing listeners (both hearing aid users and non-hearing aid users), Strelcyk et al. [15] identified the most problematic factors as:

- *'Commercials are excessively loud'*,
- *'I get annoyed when I am watching TV and other people are talking in the room'*
- *'I cannot understand what is being said because background music and sound effects are too loud'*.

Accents were also identified, being the seventh most common problem in the same closed-ended question. When asked to list what other problems the respondents encountered whilst watching TV, the most common responses were outside noise interference and quality of closed captioning. The majority (79%) of respondents indicated that the primary sound reproduction mechanism they used were the loudspeakers in their television set. Data about how often the respondent utilised the remote control was also collected, allowing the authors to show that those who reported more problems also had greater remote control usage. This led them to theorise that individuals were utilising the remote control to try and mitigate the problems they were facing.

Locating where these factors arise, and can be mitigated, within the production chain is more complex than identifying the problems themselves. The 'Audibility Problem Space', defined by Armstrong in a 2016 BBC White Paper, and reproduced in Figure 2.2, shows the possible locations within the production chain where problems can arise [21]. These range from the original performance and content capture, through production and broadcast to reproduction in the home. A similar problem space is defined by Mapp in work on the effect of television sets on intelligibility, though he adds an additional key human factor: hearing acuity [104]. This can be seen in Figure 2.3. For truly intelligible television audio all these possible points of failure need to be managed. However, many of the issues which hard of hearing listeners face, factors related to the viewer's hearing, can only be addressed within the home environment.

Defining accessibility

Accessibility cannot be an afterthought. It needs to be part of the creative process.

– Youngblood et al in "Accessible Media" [105]

Armstrong, in more recent work, differentiates the factors based on 'Viewer's Hearing' along with 'Hearing acuity' as an individual's 'media access needs' [106]. The media access

Figure 2.3 List of factors which affect the intelligibility and clarity of reproduced movie or TV sound, reproduced from Mapp [104]

Sound Capture

- Microphone response & performance
- Location & alignment of microphone
- Microphone directivity & sound rejection
- Presence, level & spectrum of background noise
- Room reverberation
- Sound reflections
- Distortion

Audio Reproduction

- Frequency response of loudspeakers
- Location of loudspeakers
- Calibration of system
- Reverberation time / reflectivity of listening / playback room
- Presence of reflecting surfaces near TV / loudspeaker(s) and listener
- Distance of listener to loudspeakers & axial position
- Background noise level
- Sound level of audio playback
- Distortion (nonlinear) of audio reproduction system

Processing & Production

- Equalisation & signal processing
- Dynamic signal processing
- De-noising
- Addition of background music & effects

Actor / Talker

- Articulation & Clarity of speech
- Modulation
- Speech Rate
- Accent
- Familiarity / fluency of language
- Voice level

Listener

- Hearing Acuity
- Attention / Alertness
- Familiarity / fluency of language

needs of an individual, or at least the majority of their needs, must be met for a piece of content to be considered accessible. There are two types of these needs: sensory needs and cognitive needs. Sensory accessibility refers to an individual's ability to perceive what is happening within a media broadcast. Cognitive accessibility refers to an individual's ability to understand, engage with, and enjoy content. Beyond establishing initial engagement and concentration on the programme, and overcoming any sensory issues, meeting cognitive needs is about: processing of the information from the content, comprehension of the content (given differing language, cultural knowledge and norms) and memory of what has occurred within the programme.

Media access needs can be permanent, temporary or situational [107]. For example, an individual with congenital hearing loss would have permanent media access needs. Temporary needs might be experienced by someone with periodic loss of hearing thresholds from glue ear whilst situational media access needs might stem from watching content on public transport in high levels of background noise. Access needs also include whether the technology required to received, interact with and view the content is accessible [108]. This may be the media player itself or the television and remote and encompasses both sensory needs, e.g. the use of braille on a remote for visually impaired users, and cognitive needs, e.g. is the interface able to be understood by a non-expert.

Models of disability

There are two lenses through which accessibility services can be viewed: the social and the medical model of disability [107]. Under the medical model, people are considered disabled by their impairments whilst under the social model, people are disabled by society and their surroundings [109]. In the latter, the onus is on society to create environments and services which allow participation, regardless of an individual's needs [110]. In the context of accessibility, an approach guided by the medical model aims to 'fix' the individual's impairment by compensating for its deficiency. Such an example would include raising certain frequencies in a broadcast so that the mix is hypothetically the same for all listeners [111]. The social model approach accepts that our inherent differences necessitate differing services in order for the narrative and meaning of the content to be conveyed effectively to all viewers. This means that individuals may experience different versions of the content (for example with subtitles, or audio description) but, through meeting the access needs of all viewers, they all can enjoy the same narrative.

2.4.3 An aside from German broadcasting

Beyond English language research, championed primarily by UK institutes and researchers, much of the published literature on broadcast accessibility for those with hearing impairments comes from Germany. Whilst potentially presenting different challenges due to idiosyncrasies of the German broadcasting landscape and language, it appears that the majority of complaints and identified issues are the same as those found in English content. In fact, it is posited by Schukrafft [112] that given the German consonant language, English is up to 6dB more comprehensible than German in a broadcast context. As such, this summary of German accessibility research could be considered a worst case scenario for English language broadcast.

Numerous opinion papers on the causes of poor intelligibility have been published, often based solely on the author's experience in broadcasting research or production [113, 112, 114, 115]. A position paper from Siegfried at Norddeutscher Rundfunk outlines ten 'golden rules' to ensure good speech intelligibility [113]. These address much of the same issues as Armstrong and Mapp adding that the increased use of faster playback can increase speech rate and deteriorate clarity. Siegfried also expands on the issues related to microphone usage suggesting that inclusion of a microphone in shot is worth it if it improves intelligibility. A 2014 paper summarising the outputs of a round-table discussion at the Tonmeistertagung [112] stresses the effect of budgetary pressures on production choices which degrade intelligibility, such as the increasing use of lapel microphones and multi-camera technology. The latter affecting intelligibility by meaning that there are no longer multiple options in post-production to select the best and clearest performance, as it is all recorded in a single take.

A 2016 paper from Krämer follows on from this, outlining the problems from source to receiver [114], in a manner similar to Mapp and Armstrong. Krämer highlights that not all stereo broadcast is created equal and intelligibility will be affected if the stereo mix is a downmix from 5.1 and not a true stereo mix. The further issues caused by down-mixed stereo mixes are highlighted in a paper by Baumgartner et al. [115]. Similarly if up-mixing is required, the process can introduce further intelligibility issues. Krämer also emphasises consumer end human factors: user-friendliness of technology and thus, ability for the end-user to correctly operate the television and its features (which is also identified in [115]). The human factors identified also included the experience, training and resources of production staff. He also points to the effect of the unregulated loudness of speech (as compared with overall loudness which is regulated under EBU R128 [95]), which is particularly stark in films [116]. The tendency for television programming to use increasingly filmic style production and mixing is also identified as a factor in ongoing complaints about intelligibility [115]. The

most recent of these papers exploring production and delivery causes of poor intelligibility is from 2018 [117]. Whilst differentiating itself through a focus on documentary and film genres, it only echoes the factors presented in earlier works.

Some papers have delved further into intelligibility issues through survey and thematic analysis of complaints made to broadcasters. In a 2014 Diploma thesis by Hildebrandt [118], an analysis was made of complaints to German public service broadcast ARD. These were recorded over the period 2011 to 2013 and 209 were selected based on their usage of keywords relating to speech intelligibility. These were then divided into five categories, with the most commonly identified being: 'music too loud' (57.9%), 'actors' speech is unclear' (18.7%), 'Language not understood' (11.4%) 'background sounds are too loud' (6.7%) and 'poor sound quality' (5.3%). These represent very similar themes to those identified in previous English language work [102]. As such, it is interesting to note unclear speech seems to be a problem which is not confined to a single language or broadcast culture. When delineated by genre (film, documentary and other), 94.8% of complaints about speech clarity referred to films. For 'music too loud' and 'background sounds too loud', Films had around half the complaints (48.8% and 50.0% respectively). Documentaries featured minimally in complaints about background sounds being too loud, though represented 24.8% of complaints about background music being too loud. The subsequent production guidelines which stemmed from this thesis [119] identify two points not mentioned in other works. The first of which is 'incomplete sentence formation', which it notes makes understanding of speech difficult. The second is the consideration of the role the music plays in the narrative suggesting that its level, and how it is mixed, should reflect that. This implies that how music should be mixed is dependent on context, rather than a strict guideline.

One of the outputs from the HHBTV4ALL project [120] was the results of a survey of 137 hard of hearing listeners, primarily hearing aid users [121–123]. 42% of respondents identified as having profound hearing loss and 26% with severe hearing loss. Whilst German definitions of hearing loss severity [123] differ to those used in this work [124], it indicates that the respondent pool would have overwhelmingly struggled to understand speech in noise, or quiet, unaided. Using closed-ended questions, respondents were asked in which genre of television programme they found the speech difficult to understand: Film was the most problematic genre (identified by 86.1% of respondents), followed by live programmes (71.3%), sports (43.5%), documentary (32.2%), news (18.3%) and other (27.0%). Respondents ranked their perceived impact of the problems with speech understanding: 98% identified mumbling as a problem to some degree (66% a large problem). This was followed by fast speech (92%), background sound (89%), speech in a reverberant environment (86%), background music (85%), sound effects (83%) and dialects & accents (74%). These issues present similar

themes to previous work [102], however, they were identified as being more problematic by more respondents. This can likely be attributed to the high level of hearing loss exhibited by participants which is not representative of the broader hearing impaired population. The solutions which would improve broadcast speech were: improved voice clarity (85% of respondents), independent control of speech/dialogue (75%), less background noise (73%), less background music (72%) and less fluctuation in volume (70%). Whilst not presenting any new findings, this corroborates other work indicating that that the problems with speech intelligibility in English broadcasting is not due to language or the unique broadcasting environment.

2.5 Accessibility in channel-based broadcasting

Channel-based broadcasting refers to the creation and transmission of programmes which have a static combination of audio and visual elements and are designed for a particular type of reproduction system. This necessitates separate assets for different reproduction systems, e.g. stereo version and 5.1 surround version of a movie soundtrack, or requires use of additional processing (e.g. a downmix algorithm) to make the content compatible with different reproduction systems. Channel-based content can be transmitted using terrestrial broadcast or streamed online. This is how the majority of content is broadcast at the time of writing.

The following section presents a chronological review of the approaches previously investigated to improve the intelligibility of broadcast speech for use with channel-based broadcast.

2.5.1 Chronological review

Many commentators on accessibility have pointed towards the transmission of a supplementary audio channel for hard of hearing listeners, often termed a 'clean audio' channel, as being an ideal accessibility solution for those with hearing impairment [125–129]. There are two possible types: a *broadcast mix* where the additional stream is transmitted by the broadcaster and a *receiver mix* which is mixed in the receiver with some capacity for user control [130]. DVB specifications describes the features of this additional channel as *audio providing improved intelligibility* [130], either through lower levels of background sound [131, 21] or an enhanced speech track [132, 25].

The earliest work in this area was conducted by Mathers in 1991 with the BBC and *Royal National Institute for the Deaf* among other partners [131]. This used audiovisual

clips with either +6dB, -6dB or unchanged background sound levels. Subjective ratings of quality on a five point Likert scale were elicited from participants which found that a -6dB reduction in the level of background noise produced only a small improvement. Mathers does highlight (with notable foresight) that without increasing the capacity in the broadcast system to transmit speech separately, improvements for those requiring large reductions in background noise would not be possible.

A stream of work conducted by NHK in Japan to develop systems to improve speech understanding on television for elderly listeners began in the mid 1990s [133, 134]. This work is focused on addressing the challenges presbycusis presents for television listening. Throughout the course of this work, which stretches well into the 21st century [135], the main problems addressed are speaking rate; for which an automatic system to slow the speaking rate was developed [133, 134, 136, 135], loudness recruitment; for which they developed a compression technology to compensate for the effects loudness recruitment has on individuals' comfortable dynamic range [136, 135], and a method for background music reduction [137–139]. Whilst the technology is designed for the spoken Japanese language, the authors suggest it should also be effective for other spoken languages. This includes the evaluation of not only the efficacy of background separation and reduction, and speech enhancement system but also user-experience evaluation of the interface [139]. The most interesting response from this study is that, when participants weren't told the purpose of the experiment, they were more likely to rate the processing as a degradation than when they were aware of the aims of the experiment. Follow-up studies which compared background suppression with and without speech enhancement showed the majority preferred background suppression only [135]. Some of this body of work explored the effects of compensating for the average threshold loss of older individuals in a television content. This was done by measuring the threshold loss of individuals, averaging it and applying compensation for this threshold loss to the test stimuli. However this was a very small study and their methodological approach was inherently flawed: as they evaluated their system with the same listeners whose thresholds they measured to define 'average threshold loss' [137]. Whilst this may be effective for these individuals, as it has specifically been compensated for their hearing, the results cannot be generalised to individuals beyond the test cohort.

A 1998 position paper by Emmett suggests a separate dialogue only mix would be the optimal solution to audio accessibility [140]. However given the impracticality of implementing this solution in the production process, he proposed a number of post-processing and spectral solutions. Shortly after this, in 1999, the DICTION project began; this project utilised the R-SPIN test to evaluate processing which removed background sounds in analogue television [132]. Both objective responses (target keywords) and subjective ratings of clarity were

elicited as part of the research. Carmichael's work indicated that whilst, at the time of the research, signal processing could make speech sound clearer, it could not objectively improve word recognition performance.

The turn of the millennium saw a transition from analogue to digital broadcasting and many researchers sought to leverage the new capabilities of the format, such as the transmission of 5.1 surround sound, to achieve improved speech understanding [25]. The Clean Audio project funded by Ofcom began in 2003, explored how to best provide a speech only audio channel given the capabilities of 5.1 surround sound. In particular, this project demonstrated that speech, reproduced from a physical central loudspeaker improved speech intelligibility compared with speech panned centrally and reproduced over a stereo pair of loudspeakers [25, 141]. This showed a quantitative improvement gained from a central loudspeaker. This non-centre channel attenuation approach to improving intelligibility using a 5.1 broadcast has been standardised by European Telecommunications Standards Institute [142] and referenced in other broadcast standards [143–145]. This method was implemented successfully in some territories, but has generally had limited impact. This is due to non-standardised use of the centre channel by many broadcasters and producers, such as including non-speech sounds.

A 2007 BBC experiment into a music-free documentary soundtrack was conducted using the red button service for 'The Nature of Britain: Secret Britain' [21]. Whilst positive feedback was received, this exercise highlighted that production of a broadcast-quality music-free soundtrack was not as simple as removing the music but required an entirely new soundtrack and significantly increased production overheads.

A special session on Hearing Enhancement at the 125th AES Convention in 2008 refocused attention in accessible audio onto speech enhancement methods [146]. In one paper from this session, Müsch [147] argues that audio processing can reduce the cognitive effort required for comprehension. His work discussed algorithms which utilised several techniques to detect the presence of speech in centre channel and to attenuate other competing sounds in the same, and other, channels. The aim of the techniques used was twofold; to decrease listener effort and, as a consequence, to improve intelligibility. Also as part of this session some results from Fraunhofer's Enhanced Digital Cinema project were reported [148]. In this work they used pattern recognition, voice activity detection and machine learning methods to enhance the speech. Subjective ratings of speech quality and general sound quality were obtained from two cohorts: one comprised of normal hearing expert listeners and one comprised of hard of hearing children. These showed that the sound quality and speech quality of their proposed method was rated comparable to unprocessed audio by the hard of

hearing cohort, whilst the experienced normal hearing listeners rated the sound quality of the unprocessed audio higher, though with comparable speech quality.

In 2008, Vickers investigated a frequency domain two to three channel up-mix approach for speech enhancement [149]. Results indicated that existing up-mixing algorithms either provided inadequate centre channel separation or produced 'watery sound' or 'musical noise' artefacts, although little perceptual evaluation was undertaken. As recently as 2015, similar centre channel speech enhancement methods have been investigated [150]. Objective evaluation using the Perceptual Evaluation of Speech Quality metric (PESQ) showed that the algorithm caused no degradation and perceptual testing showed preference for their proposed enhancement method. However validation was conducted with a small cohort of young listeners which may not have sufficient ecological validity when designing sound systems for (mainly older) people with hearing impairments.

DTV4All project [151], which began in 2008, aimed to implement a clean audio service as a separate audio stream within the DVB multiplex. This separate audio signal was either made by the dubbing engineer or derived using signal processing algorithms from the stereo or 5.1 audio. Two studies were undertaken in this project, one in Germany and one in Spain and had a combined total of 28 hearing impaired participants. These studies drew unclear conclusions about the value of this approach to clean audio, with responses varying dramatically between genre, content, processing and individual hearing abilities [152]. The overarching conclusions from the project stressed the merits of a clean audio solution, though the evidence did not directly support this standpoint.

In 2010 the BBC Vision Audibility project repeated a similar experiment to Mathers [131], providing three mixes with varying background sound levels to participants: +4dB, -4dB and unchanged [21]. This showed the greater level of background sound definitely inhibited speech understanding but less background sound did not always provide any intelligibility improvements.

DTV4ALL was followed by the HBB4ALL project in 2013 [153]. The premise of the HBB4ALL project was to exploit the capabilities of the new HBBTV 2.0 specifications to improve accessibility, reflecting a theme that revision in broadcast standards and technology can be the impetus for accessibility improvements and research. Like DTV4ALL, HBB4ALL looked towards an automatic solution for extracting a central channel and delivering 'clean audio'. This project primarily focused on the use of the 'center cut' software [154] to process stereo audio and extract centrally panned (presumed speech) content. The first study evaluated the effect on intelligibility and listening experience of attenuated left and right channels (-9dB or -15dB), with a small cohort of mostly profoundly hearing impaired individuals (12 participants with hearing aids; 10 unaided participants) [123, 122, 155, 153, 156]. The

centre channel was either taken directly from a 5.1 mix or produced using the 'center cut' algorithm, then processed with a multi-band expander and bandpass filter to enhance the speech frequencies. For those with hearing aids, the only benefit was gained when 5.1 content was used and no additional processing was applied (as in Shirley's work a decade previously [25]). For those without hearing aids, all types of processing yielded a degradation in listening. An improvement in intelligibility was only gained when the centre channel contained speech only [155].

A second round of testing was performed with 22 people without hearing aids. In this work, participants compared three different processing styles which appear to be 5.1 content with centre channel enhancement (multi-band expansion and band pass filtering) with attentuation of -9dB and -15dB of the surround channels and -9dB attentuation of the side channels with no EQ) [156, 157]. It was concluded that -15dB with equalisation offered a small improvement over the hidden reference for just over half the content samples used. However the scant implementation detail limits reproducibility.

The final part of the HHB4ALL clean audio work was four trial programs which had been processed with the -15dB with equalisation variant of the clean audio tool and then trialled by German broadcaster RBB [156]. Participants identified problematic parts of the program on a time-line, for both the original and processed versions (by separate individuals, both hearing aid and non-hearing aid users). The processed versions were evaluated by 26 participants. Whilst some segments showed improvement as a result of the processing applied, other parts of the programmes developed new issues, where there previously were no problems [156, 158, 153].

A similar study evaluating automatic extraction of centre channel content utilising an algorithm in the Goldwave DAW was undertaken in 2014 [159]. This additionally evaluated the effect of inbuilt TV speakers, external speakers and recording studio style monitors for intelligibility (via auralisations of these reproduction methods). Participants rated the recording studio monitors better than the external speakers, and external speakers better than those inbuilt into the TV set. Improvements in intelligibility from using the centre channel extraction and enhancement were shown to have a negligible effect.

Based on the work of Shirley and others, it is evident that with a clean central channel intelligibility improvements can be made [25, 155]. So as not to necessitate a 5.1 set up, a 3.0 speaker arrangement has been suggested, implemented with three speakers or utilising a directional soundbar [160]. A 2018 study demonstrated that the optimal speech to background level is affected by overall playback level; the lower playback level likely in home scenarios requires higher dialogue levels [160]. This research has merit, due to the increasing use of soundbars. However the efficacy of this approach, like other methodologies focused on the

centre channel, relies on decisions made in the production process about how the centre channel is used.

2.5.2 Discussion

The clean audio approach, which automatically derives the speech signal from incoming, unknown stereo or 5.1 content, can be effective. Work here has highlighted though that this efficacy is highly reliant on 5.1 with speech content only in the centre channel [25, 155]. For stereo content, blind source separation still proves a challenge, though algorithms are improving [161]. The main problem with 'clean audio' still remains though – the assumption that the delivery of a clean speech signal is all that is required to provide accessible audio.

Research in channel-based broadcasting accessibility has provided a number of useful lessons about accessible audio and has served to increase awareness of the sensory media access needs of hard of hearing listeners. Much of this work, in particular strategies which rely on a clean centre channel [25, 155, 160], highlight the key point that for accessibility strategies to work, they must be considered in the production stage; either by delivering alternate content or producing content in a consistent, standardised way. Conversely, if accessibility strategies add significant overhead to production, they may not be adopted by broadcasters [21]. Speech enhancement approaches continue to have little traction in linear broadcasting and where they do demonstrate usefulness it is not in improved intelligibility, but in reducing listening effort [147, 103]. Finally, given the variability in hearing loss, it is unsurprising that many accessibility studies investigating optimal balances fail to draw strong conclusions [131, 21]: it is evident that a more personalised approach is needed.

2.6 Personalisation for accessibility

In channel-based broadcasting, the required technology to easily deliver separate dialogue or personalisable content does not exist. This, coupled with the variability in hearing impairment, means 'clean audio' research has had minimal impact on audiences' experience of TV audio. However, the development of object-based media technology gives a foundation for personalised content which can improve accessibility. This section outlines the technology which makes such personalisation possible and explores the most useful ways in which personalisation could improve the viewing experience for hard of hearing listeners.

2.6.1 Object-based broadcast for accessibility

The next generation of audio broadcast technology is called 'object-based broadcast' or 'object-based audio'. In object-based broadcasting, different broadcast elements including speech and non-speech audio elements, as well as other elements like the visuals, can be treated as independent 'objects'. These objects, transmitted along with metadata, remain separate until they are rendered at point of service based on their associated metadata [23, 22]. This allows for the delivery of personalisable [162, 163], immersive [164], non-linear [165], responsive [166] and interactive content [106, 167, 24]. A visual representation of the difference between object-based and traditional channel-based broadcasting methods can be seen in Figure 2.4.

Object-based audio formats for efficient distribution include Dolby Atmos, MPEG-H and DTS:X. These have reached differing levels of maturity and have differing personalisation capabilities. At the time of writing, object-based audio technology has been rolled out in several territories. Initial broadcasts have been limited in their exploitation of the personalisation potential of object-based audio [19, 20]. MPEG-H is being broadcast in South Korea as the sole audio codec for the country's terrestrial UHDTV broadcasting system [19] and includes facility for audio description and dialogue to be broadcast as separate, personalisable objects [168]. 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 [169].

Object-based audio opens up the possibilities of personalising audio presentation based on individual viewer preferences, sensory or environmental needs. By facilitating this personalisation, it has the ability to improve the accessibility of broadcast content in a way not possible in channel-based broadcasting [170, 171]. However, it would not be practical for the broadcaster, nor enjoyable for the end-user, if all possible parameters of the broadcast are personalisable. To maximise likelihood of adoption and greatest benefit for hard of hearing audiences, an understanding of the possible personalisation parameters and the work that has been done to explore them, is required.

2.6.2 Dimensions of Personalisation

This section outlines the possible personalisable dimensions of object-based audio which may be of benefit to hard of hearing listeners through a conceptual analysis of broadcast speech intelligibility literature. This resulted in the identification of three main dimensions of personalisation: speech to noise ratio, spatial separation and redundancy. These dimensions,

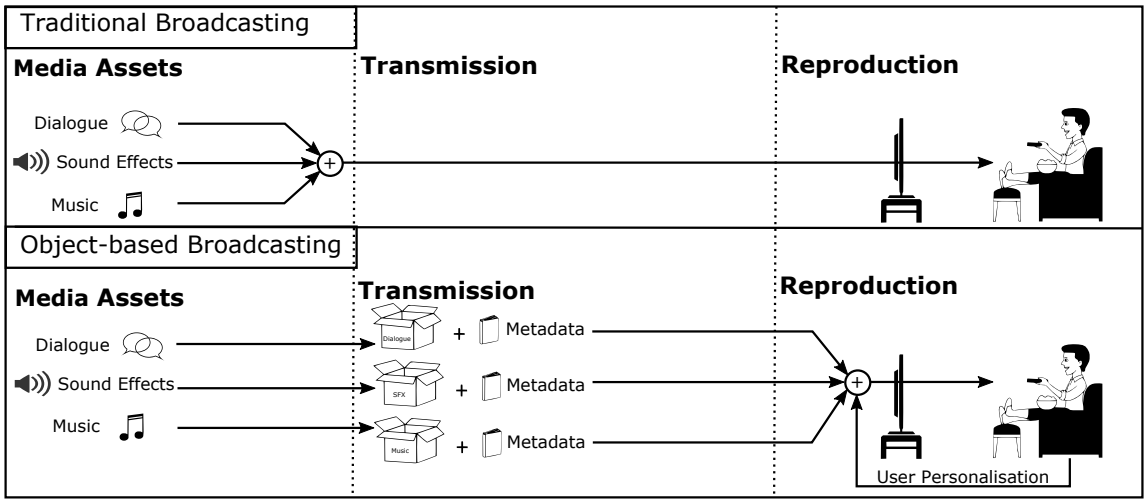


Figure 2.4 Visualisation of object-based broadcast in comparison to channel-based, noting the potential end-user input which can influence reproduction of the media assets in the home.

and previous work exploring how altering these parameters may affect broadcast speech understanding, are outlined below.

Speech to background ratio

The first response when speech is not understood on television is to turn up the volume [172, 15]. For a small proportion of those with hearing loss additional volume is sought through the use of headphones, hearing aid induction loop systems or TV to hearing aid streamers [15]. 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 [173]. When the *attenuation* caused by hearing loss is overcome, the *distortion* loss, or effective speech to noise reduction, remains [38].

Improving signal-based speech intelligibility requires an increase in the speech's level relative to other sounds to ensure that more of the original acoustic speech signal can be hear. A number of additional studies to identify optimal speech to background ratios have been carried out (many as part of the studies described in Section 2.5). Mather's work [131] suggested that a 6dB reduction to the background sounds produces a small improvement in quality. The BBC audibility project [21] showed that a 4dB increase in background sound, compared to the speech, degraded speech understanding. As part of a 2018 study developing an objective intelligibility metric for broadcast content, a study of listener's preferred speech to background ratios for five types of speech only content was made [174]. This showed that preferred speech to background ratios are in the range -6.1 to 15.6 dB, and vary significantly

between content types. The largest loudness difference was for a piece of content with speech over speech content and the highest level of background sound was set for a radio drama.

Given the limited number of English language studies into broadcast speech to background ratios, this section will also outline results from German and Japanese language studies. Whilst the ratios found in these studies should not be expected to be directly applicable to English language broadcasting, it demonstrates the variability of ratios found both within a single language and across languages.

For German language content, a 2013 study investigated the speech to background ratios for film, documentary, sport and magazine style content with a variety of normal and hard of hearing listeners [118]. Three speech to background ratios were trialled: -2dB, -7dB and -10dB (referenced to speech at 0dB). The -7dB speech to background ratio received the most positive responses from participants, with respect to enjoyment and listening effort.

Two additional German language studies have explored the optimal speech to background level when the speech is spatially separated from the remaining audio content. Work in the HBB4ALL project showed that for unaided listeners with high levels of hearing impairment, ratios of background sound of -9dB and -15dB, compared to 0dB speech both provided intelligibility improvements [155, 156]. A 2018 study explored the required speech to background ratio of 3.0 audio on soundbars and external speakers at different reproduction levels across different content with normal hearing listeners [160]. This found that at higher reproduction levels, 70dB SPL, up to +1.3dB ratio of background sound to speech was acceptable (for speakers and sound bar) whilst at lower reproduction levels, 50dB, necessitated a ratio in the range -2.4dB (speakers) to -3.0dB (soundbar).

Studies conducted by NHK, for the Japanese spoken language, have investigated balance between foreground and background sound [175]. In documentary content the ratio of background to speech was set by mix engineers as -9 $\pm$ 3 dB. In musical programming and sport, ratios close to 0dB or where the music was louder than the speech were chosen by the same mix engineers. However, they used a variation of the ITU BS.1770 loudness standard, which adopts smaller time constants for average programme loudness and the extent to which it is comparable to other loudness measures is unknown. Further studies by NHK into how elderly listeners perceive overall programme loudness showed that listeners found the background to be too loud when the difference between narration and background sounds was not at least 6 phon [2] [139]. It also found that the background was perceived as too loud when background sounds are more than 2.5 phon louder than the average narration level in periods of content where only background music or sounds are present. Follow up

[2] Phons are a unit of loudness designed to incorporate human's differing loudness perception at different frequencies. *The 'loudness level' of a sound is defined as the intensity in dB SPL of that 1 kHz pure tone which sounds equally loud*[176].

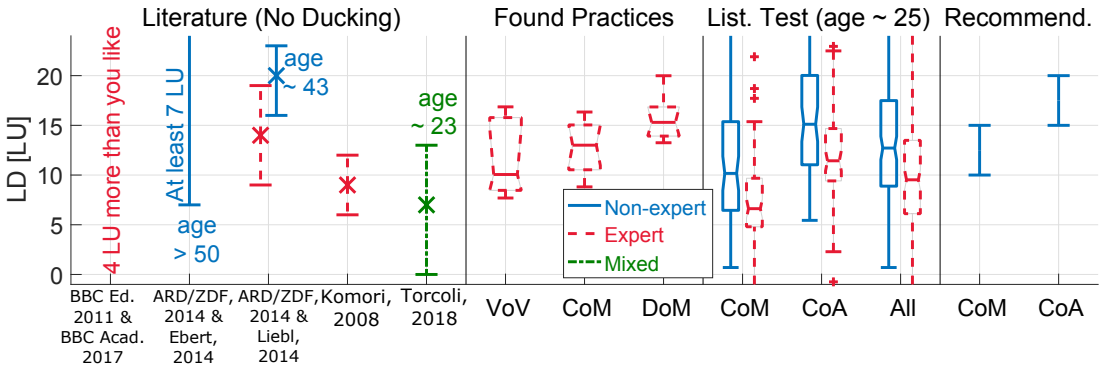
studies trialled a prototype system with a background sound suppression technology with 11 individuals over sixty years of age with mild-moderate hearing loss [135]. Responses from this group fell into two categories: those who preferred louder background; on average 1.3dB louder than the original background balance and those who preferred a softer background; on average -2.79dB less than the original background to speech balance. However the study does not outline the original balances, which limit the interpretibility of the results.

Other studies by NHK asked 12 mixing engineers to set what they felt were the upper, lower and 'best' levels of background sound to speech ratios in 30 second excerpts of documentary, music, sport and drama content [175]. The 'best' ranged from +2.9dB to -2.8dB, with the highest upper limit at +7.2dB. No background sound in a program was set on average lower than -7.1dB. Whilst not a generalisable result, this highlighted that mix engineers in Japanese broadcasting perceive the acceptable limit of background sound to be much higher than the public do.

When both the background and foreground sounds are speech, termed voice over voice, a larger ratio is required (as is the case when translated speech is placed on top of ducked original speech). A 2014 German language study found original speech was set in the range of -8.9 and -18.8 dB below the new speech, with the upper and lower quartile values for the participants' preferred levels being in the range -14.9 and -21.4 dB [177]. The results of this are seen in the literature section of Figure 2.5. The specifications for this type of content from German broadcaster ARD recommend a loudness difference of -20 dB for voice over voice passages [96]. This work concluded that the preferred level varied greatly between participants and the different content examples. More recent work from Torcoli et al. surveyed the current ratios at which speech and other content is ducked below the main dialogue [178]. A cohort of expert listeners and a second cohort of naive listeners then investigated the preferred ratios. Their results are summarised in the third and fourth sections of Figure 2.5. A key result from this work is the preference of expert listeners for the background to be on average 4dB louder than the naive listeners. This is similar to the result from Japanese mix engineers who, as also expert listeners, preferred much higher average background sound levels to naive listeners. Additionally Torcoli et al.'s work shows that the range of ratios used for voice on voice content can be extremely small. Other work also from Torcoli et al., demonstrated that for young normal hearing listeners, preferred level differences between the speech and background sound ranged from 0dB to 13dB and was a highly personal choice [179, 180].

Outside of broadcast research, a 2018 study by Wu utilised recordings of twenty older adults' day-to-day acoustic environment to explore the speech to noise ratios experienced in the real world scenarios [181]. This found that as the noise levels increased from 40 to

Figure 2.5 Integrated Loudness Difference (LD) between foreground speech and background for the default mix in TV audio. Visual summary of literature review, common practices found in the analyzed documentary programs, results from the preference listening test in [178], and final recommendations for Commentary over Music (CoM) and Commentary over Ambience (CoA). A modified loudness measurement is used in [175] and it is unknown to what extend it is comparable with the other values. Image reproduced from Torcoli et al.[178]



74 dBA, speech levels systematically increased from 60 to 74 dBA. The resultant speech to background ratio reduced from -20dB to 0dB, with the majority of measured ratios in the range -2 and -14 dB (62.9%). Very noisy situations that had speech to noise ratios above 0 dB comprised 7.5% of the listening situations. Broadcast studies have indicated that preferred background to speech ratios range from between -15dB and -2.4dB and can be as high as +1.3dB for high reproduction levels (70dBA) [131, 21, 118, 155, 156, 160]. These match closely with the range of speech to background ratios found in day to day scenarios by Wu [181]. Given that day-to-day scenarios present differing levels of difficulty for understanding, depending on the individual, it is unsurprising that no consensus on a 'correct' speech to background ratio has been found between and even within these broadcast studies. Furthermore, as will be outlined in Section 4.3, the speech reception threshold is itself a feature used to characterise the degree of an individual's disablement from hearing loss.

Given that the level of speech and other elements can be separately controllable in object-based audio, speech to background ratio becomes an easily personalisable dimension. Given that variety in preferred ratios, it has potential to be of significant benefit as personalisable parameter for hearing impaired listeners.

Spatial separation

Spatial release from masking refers to the increase in intelligibility which is gained when the target speech is spatially separated from the noise source [81, 182]. When sounds are spatially separated different acoustic cues are imposed on the speech and competing sounds based on their location. These differing acoustic cues aid the segregation of the two audio streams in the brain, resulting in the speech being more easily understood. This is a facet of complementary intelligibility. Spatial separation yields the most benefit when the target and masker are easily confused, such as two competing voices. An advantage is also gained when the target is directly in front, due to binaural summation: as the target signal is received by both ears and summed. Normal hearing listeners gain a benefit of up to 12dB as compared with a co-located source [182]. A useful review of the concepts and research in spatial release from masking can be found in [182].

However not all listeners will benefit from spatial separation to an equal degree. It has been long demonstrated that some *'pinnae, in their role of transforming the spectra of the sound field, provide more adequate (positional) cues than do others'* [183]. Given this variation, spatial separation will provide less benefit for some listeners. This extends to hearing impaired listeners who benefit to a reduced degree [184], dependent on their specific hearing impairment and localisation ability.

In broadcast, the benefit gained from spatially separating and centrally locating the speech has been explored in a number of projects, either tangentially [155, 149] or directly [25, 160]. The Clean Audio project, described in Section 2.5.1, showed that a central loudspeaker improved intelligibility over a phantom centre [141, 25]. It can also be seen from Goossens work on 3.0 audio that when a central channel is employed, much smaller speech to background ratios are acceptable [160]. This includes ratios where the background sound is louder than the speech.

Object-based audio, where the location of the speech object can remain separately controllable, gives the potential for the spatial location of the speech to be a personalisable dimension. In particular, personalisation of the speech's location may be useful to place the speech directly in front of the listener when, due to the listening environments, a standard central channel would not be in front of the listener. Additionally, given the developments in flat panel television technology which utilise the television screen itself as a sound transmitter, allowing objects to be rendered coherently with their spatial location on the screen [185].

Redundancy

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

– Holman in "Sound for film and television" [186]

Redundancy, or complementary cues which may be superfluous for normal hearing listeners in quiet, can facilitate understanding in less favourable conditions or for people with hearing loss [78]. This is particularly relevant to audio-visual media as it does not represent a standard speech in noise problem [187, 15]. Non-speech broadcast content includes music, on-set effects, recorded 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 subtitles (also known as captioning) 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'* [188].

The importance of redundancy in understanding speech has long been understood in terms of how context, word familiarity and syntactic structure are used [189, 82, 77]. 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 recognising 37% of keywords up to 76%) [82]. Recent adaptations of his work consistently demonstrate this effect [190–192].

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 better when spoken by a familiar voice, spouse or close friend, than by a stranger [193]. Even for speech which is previously unfamiliar, familiarising a listener with the speaker's voice beforehand can result in intelligibility gains [84]. Another type of single mode cue is non-speech sounds. A 2016 study by Hodoshima showed that some types of preceding sounds aid intelligibility of urgent public address style speech [194].

Multi-modal cues have been investigated in many studies, including the interaction of different complementary intelligibility cues [86, 195, 85]. 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 [86]. It showed that by age five children can utilise the situational context of speech. Zekveld et al. analysed 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 [195]. Spehar et al. have investigated the effects of different types of contextual cues showing that participants benefited from both visual and speech-based context [85]. 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 dialogue for both speech clarity and enjoyment ratings [187]. 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 personalisable, even with the aid of object-based audio. Semantic context is already present in most dramatic content, providing built-in redundancy. Through an ability to control the reproduction levels of different objects, or groups of objects, object-based audio presents an ability to personalise the level of redundant non-speech sound cues. More broadly, object-based media broadcast presents the potential to personalise 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.

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 utilise complementary intelligibility cues [147]. 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 [196]. This work showed that for instructional messages, additional audio elements can overload the listeners' working memory. 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 [197]. A 2010 study by Aramaki et al. showed that categorisation of ambiguous non-speech sounds takes longer than for typical sounds [198]. These works suggest that the amount of redundant information and how it is presented requires personalisation to ensure that an optimal balance between improved intelligibility and cognitive needs is maintained.

2.7 Systematic review of previous work

Since the advent of object-based broadcasting, there have been many investigations to determine how the technology may be used to improve the accessibility of broadcast content

for hard of hearing individuals. This section reviews all relevant projects, in light of the dimensions of personalisation derived in Section 2.6.2. The aim of this is to determine which approaches have shown the greatest efficacy and whether there are potentially beneficial strategies which have not been explored.

To ensure that all relevant developments in personalised and accessible object-based audio were identified, a systematic review methodology was employed. This review considers the literature in terms of projects due to the evolving and active nature of the research area, similarly to [21]. 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 30th July 2019 were considered. Outputs from this doctoral work were not included.

For inclusion, project aims had to meet the following criteria:

- to enable personalisation of an element of broadcast audio, which the end-user has control of at time of consumption
  AND
- use object-based audio to do so
  OR
- rely on object-based audio methods for eventual implementation of a theoretical investigation.

18 projects meeting this criteria were identified. Two additional projects were initially identified and later excluded as they did not specifically describe *audio* personalisation [167, 199]. The majority of projects addressed speech to noise ratio in some manner (16 projects). 7 projects explored some manner of spatial separation whilst only 4 projects explored personalisation of redundancy.

The following sections outline these projects. First those projects which provide general audio personalistion capability and then those projects which present personalisation specifically designed for accessibility. Projets defined as 'accessibility projects' are those which in their public outputs specifically indicate either that the aim of their work is for access service applications or that their work is targeted towards viewers with impaired sensory perception. Their characteristics and the organisations involved in each project are outlined in Tables 2.2 and 2.3 respectively.

2.7.1 Object-based audio personalisation

Prior to the first broadcast of object-based audio formats work using the Web Audio API and SAOC-DE had been carried out to explore how personalisation could be employed. The BBC

Table 2.2 Projects used in systematic review including whether personalisation is accessible (Access.), personalisation dimensions (Speech to background ratios (SBR), Redundancy (Red.) or Spatial Separation (Spat.), project dates and references.
✓* indicates visually impaired only, †indicates work completed as part of this thesis (and subsequently omitted from the review).

	Title	Began	End	Audio	Access.	SBR	Red.	Spat.	Refs
1	FascinatE Project [200]	2010	2013	✓	✗	✓	✓	✓	[187, 201, 202]
2	Netmix	2011	–	✓	✗	✓	✗	✗	[162, 203]
3	Audibility in Radio	2012	–	✓	✗	✓	✗	✗	[162]
4	BBC Football	2013	–	✓	✗	✓	✗	✗	[163]
5	Orpheus Project [204]	2015	2018	✓	✗	✓	✓	✗	[205, 206, 24]
6	Venue Explorer/Immersive Coverage of Spatially Outspread Live Events (ICoSOLE) Project [207]	2013	2016	✓	✗	✗	✗	✓	[208–210]
7	2-Immerse Project [211]	2015	2018	✓	✓*	✓	✗	✓	[212, 213]
8	‘Adaptive, Personalised “in browser” Audio Compression’	2015	–	✓	✗	✓	✗	✗	[214]
9	‘Delivering Object-Based 3D Audio Using The Web Audio API And The ADM’	2015	–	✓	✗	✓	✗	✓	[215]
10	‘Exploring object-based content adaptation for mobile audio’	2016	2018	✓	✗	✓	✗	✗	[216–218]
11	S3A: Future spatial audio for the home [219]	2014	2019	✓	✗	✓	†	✓	[174, 220–228]
12	3D Soundbar Audio Personalisation	2014	Pres.	✓	✓	✓	†	✓	[229–231]
13	DTS Dialogue Enhancement	2015	–	✓	✓	✓	✗	✗	[232]
14	SAOC-DE and MPEG-H Dialogue Enhancement	2011	Pres.	✓	✓	✓	✗	✗	[233–235, 180, 20, 161]
15	‘Personalized object-based audio for hearing impaired TV viewers’	2015	2017	✓	✓	✓	✓	✗	[170]
16	‘Adapting Audio Mixes for Hearing Impairments’	2017	–	✓	✓	✓	✗	✗	[111]
17	Immersive Accessibility (ImAc) Project [236]	2017	Pres.	✓	✓*	✗	✗	✓	[237, 238]
18	‘Determining appropriate categorical boundaries for audio objects with regards to their importance to narrative clarity’	2017	2017	✓	✓	✓	✓	✗	[239]
					Totals	16	4	7	

Table 2.3 Corresponding project partners for each of the projects in Table 2.2.

	Organisations
1	BBC R&D, University of Salford, TNO, Fraunhofer, Technicolor, Interactive Institute, ARRI, Joanneum Research Digital, Softeco Sismat, UPC, Alcatel-Lucent
2	BBC R&D, Fraunhofer IIS
3	Swedish Radio
4	BBC R&D , 5Live Sport
5	Fraunhofer IIS, EURSCOM, BBC R&D, IRT, Elephantcandy, Trinnov Audio SA, B<>com, ircam, Bayerischer Rundfunk, Magix Software
6	Joanneum Research, Technicolor, Vlaamse Radio- en Televisieomroeporganisatie, iMinds, bitmovin, BBC, Tools at Work Hard+Soft
7	BBC R&D, BT, CWI, Illuminations, IRT, Cisco, Chyron Hego
8	BBC R&D
9	BBC R&D
10	BBC R&D
11	University of Salford, University of Surrey, University of Southampton, BBC R&D
12	University of Southampton, AudioScenic
13	DTS
14	Fraunhofer IIS
15	DTS, University of Salford
16	Queen Mary University London, Queens University Belfast
17	I2CAT, RBB, University of Salford, Corporació Catalana de Mitjans Audiovisuals, Motion Spell, Universitat Autònoma de Barcelona, Anglatècnic, IRT, RNIB
18	University of Salford

All project partners are listed, not all listed partners participated in the audio personalisation workstreams reviewed in this work

and Fraunhofer IIS carried out 'Netmix' in 2011, an experiment using a live broadcast of the Wimbledon Tennis Championships which allowed end-users to select between seven options for relative level between commentary and court ambience [162, 203]. Two distinct patterns were apparent in listener's 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, yielding similar results [162].

Other work utilising the Web Audio API has been completed by the BBC. Personalised dynamic range control, based on the end-user's preferences, needs and listening environment, was proposed and informally evaluated [214]. A demonstration of object-based audio and the audio definition model using the Web Audio API proposed the ability for the user to mute and unmute individual objects as well as select binaural or stereo rendering in the browser [215].

Further live broadcast experiments by the BBC with football content allowed viewers to customise which team's end the crowd noise came from [163]. This work utilised three audio streams: on pitch sounds, commentary and crowd noise streams. Users could freely adjust the crowd ambience and commentary level on a scale and could also adjust between home and away crowd sound. 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 [201, 202]. 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 personalise levels of on-pitch, crowd and commentary sounds during a game [187], although formal user-testing of this was not carried out.

The work from the FascinatE project [200] was further developed into a BBC project called Venue Explorer and a linked project; the Immersive Coverge of Spatially Outspread Live Events (ICoSOLE) Project [207–210]. This project explored the adaption of object-based audio based on the location in video scene on television or spatially within a virtual environment. Outputs from these works included the Glasgow Commonwealth games, where users could move around the athletics stadium and receive the audio of a particular event [208], and Edinburgh Fringe Festival [210].

Automatic selection of a background level to optimise intelligibility has been investigated by Tang et al. [174]. The system utilises an objective intelligibility metric to analyse the intelligibility of the speech. If the intelligibility falls below a predefined threshold, the system exploits the separation between speech and other the sounds in object-based audio, to adjust

the overall speech to background ratio. The threshold at which adjustment occurs could be personalised.

Personalisation to improve the listening experience in non-ideal environments and mobile devices has been explored by Walton et al. [217, 216] by allowing users to adjust the foreground/background balance. Background sounds were defined as diffuse and ambient sounds, foreground sounds as dialogue and prominent sound effects. Environmental noise had a significant effect on the mix preferences of participants. Again, the results highlighted two distinct clusters of behaviour: 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.

The 2-Immerse project [211] developed companion screen technology which, among other features, could control some of the reproduction parameters of the main screen content [212]. They undertook a trial with the MotoGP where viewers could personalise visual aspects, such as the video content they viewed, how the leaderboard was presented and were also given independent control over the level of the commentary and the ambient audio. Users scored nearly all features, including the audio personalisation, as 9/10.

As part of the S3A project [219], Demonte et al. also investigated personalisation potential for mobile and small screen devices [220]. This work explored the effect on intelligibility of binaural auralisation of noise, speech or both, as well as the effect of visual information. The study utilised the GRID audio-visual corpus [240] 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 externalised to the screen and masker noise reproduced in stereo in the headphones. This effect has been attributed to binaural release from masking [241] and audio-visual coherency when the speech appears to be coming from the speaker on the screen. The effects of binaural stream segregation and the availability of glimpses in the better ear also likely contributed [242]. This may be due, in part, to binaural signal processing increasing the relative gain of the frequency range containing consonants in speech.

Early work in the S3A project demonstrated a 3D audio radio drama, which was streamed in real time from Salford to London and rendered locally on an array of 17 loudspeakers and two subwoofers [221]. The mix could be modified in the reproduction environment remotely through a graphical web interface.

More recent work on the S3A project has leveraged object-based audio to create immersive audio scenes in the home without the need for 5.1 or other surround sound system [222–226, 228]. This approach, called media device orchestration (MDO) utilises ad hoc arrays of personal and mobile devices along with additional metadata which directs the audio

objects to the appropriate connected device. Objects can be directed based on the location of the device, the other audio objects which are already playing or to provide additional hidden, or *easter egg*, content. This ability enables MDO to augment a sound scene for improved immersion and also has potential for specific objects, e.g. narration, to be sent to a specific individual's device. This has currently unexplored applications in accessibility and intelligibility improvement, by manually (through device location) improving experience for this system has been evaluated qualitatively [222] and quantitatively [225, 223] and shown to provide comparable quality of experience to 5.1. The efficacy of MDO has been further explored with a production trial, where a bespoke piece of content was created and then publicly trialled [224].

The Orpheus Project [204] has explored the user-experience of object-based media [205, 24]. Early in a project two workshops were undertaken with broadcasting professionals, to identify the desirable features and requirements of object-based production and delivery. From this three key dimensions of experience were identified: audio (personalisation, intelligibility and immersion), information (contextual metadata of content and accessibility) and usability (human interaction and user interface). 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. The project then developed a number of prototype technology and production tools for creating and consuming object-based audio content. These include implementation of the foreground/background balance on mobile, PC and AV receiver platforms as well as production tools which allow for monitoring of the object-based personalisation features, including the foreground-background balance and a 'clarity' button that adjusts dynamic range [206, 24]. A user-experience study of these features was also undertaken with a large cohort 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, with one participant stating: *"Background noise no longer impairs listening pleasure"*. 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.

2.7.2 Personalisation for improving accessibility

Early work to address the needs of hard of hearing listeners was conducted by Fraunhofer IIS and termed Spatial Audio Object Coding for Dialogue Enhancement (SAOC-DE) [233, 234]. SAOC-DE was designed to complement existing 5.1 and stereo broadcast systems and transmitted un-mixing metadata which could separate audio objects from the audio mix [233]. This approach can be considered as 'informed' source separation. In intelligibility tests using the Oldenburg Sentence test [243] and applause style background noise, it was demonstrated that SAOC-DE improved sentence recognition accuracy from 34% to 81%. Other dialogue enhancement work has been completed, aimed at archival content where original stems or objects are not available. This approach integrates multiple blind source separation techniques to optimise the extraction and enhancement of the dialogue [235, 161]. The MPEG-H format is then used to facilitate end-user personalisation of the speech level [179, 20, 161]. In a recent evaluation with a small cohort of German speaking experienced listeners, participants were asked to adjust the speech level to a point which balanced quality and dialogue intelligibility [161]. The dialogue objects were derived either from object-based MPEG-H content or using source separation algorithms from archival content. This showed that participants manipulated the audio objects derived from the source separation algorithm in a similar manner to audio objects which were produced as object-based content.

DTS have also presented a dialog-based personalisation solution [232]. The proposed algorithm specifies dialogue control and enhancement. Alongside this is a protection mechanism to ensure appropriate levels of dialogue compared to other program content are maintained through sections where levels change substantially. The algorithm makes use of object loudness metadata.

As part of the 2-Immerse project [211], a companion screen technology which can deliver an audio description track from a synchronised device has been developed [213]. This allows viewers with sight loss to have some control over the audio description, and forms the basis of delivery for other types of personalised object-based audio to an individual listener. The most likely implementation for this technology would be over headphones, which can isolate the listeners from the communal experience of watching television.

A solution to providing individualised audio whilst maintaining the communal experience has been proposed by Simon Galvez et al. [229] and developed further as part of the S3A project [230]. This solution utilises highly directional beam-forming, implemented in a consumer-style soundbar, to deliver personalised audio to only the listener requiring it whilst providing a standard audio mix, or a second personalised mix, to additional listeners. This is also adaptive to the listener's position, through the use of video tracking to ensure the individualised reproduction follows the listener [231]. A demonstrator has been developed

which allows two listeners to separately make adjustments to the mix they are each receiving, whilst sitting next to each other.

It has been highlighted that the potentially large number of audio objects in a television program, and the fact that object-based audio 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 [21]. Work by Woodcock et al. [227] has investigated how people cognitively categorise 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 categorisation scheme has been utilised in a reduced form in subsequent work where users were given control of four sound categories: dialog, music, foreground effects and background effects [170]. This project, a collaboration between DTS and the University of Salford, presented hard of hearing participants with an interface allowing them to adjust the level 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.

Other work, by Pearson, has explored the effect of sound categorisation for object-based content on overall narrative clarity [239]. In this work, participants listened to three different mixes of four audio-only scenes in background noise and rated the narrative clarity of each scene out of 100. The three mixes were: *all*, which included all the audio objects, *dialogue only* with no non-speech sounds and *dialogue + key FX* which had non-speech sounds which were key to the narrative as well as the dialogue. For both the radio drama scenes, the narrative clarity of scenes with a dialogue only mix were rated significantly higher than the other two mixes (*all* and *dialogue + key FX*), which were not significantly different. For one of the short film scenes, there was no significant difference between any of the mixes and for the other, both the *all* and *dialogue + key FX* mix were rated much higher than the dialogue only. This study suggests that for young normal hearing listeners, in high levels of background noise, key FX mixes and the broadcast mix have similar narrative clarity and, depending on the content, a dialogue only mix may improve or degrade clarity. However the generalisability of these results to content beyond the two studied is unknown.

The Immersive Accessibility project (ImAc) [236] is in the early stages of exploring accessibility for 360 degree content [237]. The project focuses on audio description, signing and subtitling. Initial focus groups have suggested broader accessibility applications as well [238]. Whilst the focus groups were extremely limited, it was noted that a desirable feature of the audio description was for the foreground, i.e. the current main field of view, to be increased in level whilst the background audio (not currently in view) is attenuated.

2.7.3 Discussion

Object-based audio 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 personalisation for consumers and the industry are outlined.

Speech to background ratio

Speech to background ratio is the most commonly implemented dimension [162, 216, 232, 174, 187, 203, 205]. It is clearly a desirable and effective personalisation parameter, given the balances chosen by even normal hearing listeners. It is also powerful in that it can be leveraged to improve intelligibility [234], combat adverse listening environments [217, 216, 205] and increase immersion [162].

Object-based strategies highlight an important question; how to offer personalisation for legacy and other pre-mixed content. Hybrid strategies, like SAOC-DE which make use of the increasing efficacy of source separation algorithms [244, 245] and the personalisation capabilities of object-based audio, 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 [232].

Personalised dynamic range control and compression [214] has potential to intelligently interact with hearing aids to prevent problems deriving from multi-band compression being applied multiple times [15]. This is an unexplored area which has significant potential as the ability for devices to communicate directly to hearing aids improves. Whilst the exploitation of hearing aid technology to provide personalisation 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 [216, 217, 212, 205] or a selection between

multiple predefined mixes [163]. Automatic adjustment shows potential to lower the barrier to accessing such a feature further, particularly if the intelligibility threshold can be set for individual listeners, listening scenarios or preferences [174]. Whilst significant work has explored personalisation of speech to noise ratio, automatic adjustment represents an unexploited area.

Whilst conceptually simple for the end-user, this simplicity relies on utilising the target speech vs. masker (everything else) paradigm [233, 162] or ad hoc definitions of foreground and background sound [217, 216]. However, the distinction between useful and masking sounds is more complex. This is evidenced by the personal preferences reported by Shirley et al. [170]. Whilst Woodcock et al. demonstrate some generality in people's categorisation of broadcast sounds [227], the potential for each group of sounds to act as a masker is not generalisable.

Spatial separation

This dimension has received comparatively limited research attention and may be in part due to the majority of terrestrial broadcast utilising 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 personalisation through binaural rendering [215, 205, 220] or sending alternate mixes to secondary devices [222, 213]. The use of binaural rendering has the additional advantage of providing useful location cues to listeners who may also have some sight loss [246]. Binaural reproduction over soundbar technology has potential to provide personalised speech to noise ratios [230] and also increased spatial separation for individual listeners [229]. As highlighted by Demonte et al., the effect of binaural auralisation on the intelligibility of speech is not well known and further research is required [220].

Beyond the technological challenges however, spatial separation presents a parameter which is conceptually difficult to personalise. 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. Furthermore, investigations to determine whether consistent benefit is gained across audiences, particularly those with hearing loss for whom localisation abilities are often reduced, would be needed. Such a control could take the form of a 'spread out' button, personalising levels of spatial separation.

Redundancy

Multi-modal redundancy cues, in the form of subtitles, have reached near ubiquity as an access service in some regions. However, personalisation of other redundancy cues, particularly single mode cues, has seen less exploitation. Multi-modal informational redundancy provided by transcripts [205] 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 [199]. Similarly, a more recent pilot from BBC R&D developed a prototype podcast player which provides transcripts, charts and images referenced in the podcast and related links [247]. These kind of approaches, coupled with tools for synchronised second screen content [248], offer technology to deliver supporting information. This could be leveraged to provide additional audio that would allow users to familiarise 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 personalised and exploited [170]. However, as highlighted in Section 2.7.3, the challenge is then how to categorise these sounds in terms of their relative usefulness and masking potential whilst making personalisation easily accessible to the end-user. Follow up work by Pearson contributes to our understanding of how sounds important to narrative clarity (rather than speech clarity specifically) may affect overall quality judgements [239]. However, significantly more research particularly under controlled conditions is required.

2.7.4 Conclusions from the systematic review

It is evident that 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 Paulus et al. [161] can be facilitated by object-based audio. Speech separation algorithms such as these could also be informed by intelligibility metering and adaptation as described in [174] to automate for optimum intelligibility.

For object-based audio methods which allow personalisation of the speech to noise ratio, significant evidence of the efficacy of this feature and desire for it have been shown [205, 216, 170, 212, 217, 233, 162]. The focus of future work should be on how this can be

implemented in current and future object-based audio formats, as well as developing simple but effective interfaces.

Much unexploited potential still exists in the areas of spatial separation and redundancy. Spatial separation can be seen to given significant advantage with respect to intelligibility, however its implementation presents a conceptual challenge. It is likely that the most benefit for the largest audience from this dimension will be achieved through the provision of binaural mixes. Whilst significant implementation of multi-modal redundancy (in the form of subtitles) exists, comparatively minimal explorations have been made about how components of the acoustic and visual scene in broadcast content can improve understanding and access. Furthermore, a better understanding of how the different audio elements interact with speech understanding would help inform the implementation of speech to background ratio personalisation in a more nuanced manner than the foreground/background approach.

2.8 Chapter summary

This chapter has outlined the concepts core to the study of broadcast accessibility for hard of hearing listeners, including intelligibility and how hearing loss is characterised. The challenges currently faced by hearing impaired individuals accessing broadcast content were summarised and followed by a chronological review of research in accessible broadcast audio. Personalisation of broadcast content, and the technology to support it, were outlined and the chapter concludes with a systematic review of current research activity in object-based audio personalisation.

The key conclusions which can be drawn from this chapter are:

- Hearing loss is characterised by both threshold loss, *attenuation* and suprathreshold loss, *distortion*, which results in each individual's hearing loss being quite unique.
- Overall intelligibility is influenced by both the ability to receive a speech signal (signal-dependent intelligibility) and utilise additional non-speech information to decode it (complementary intelligibility).
- Object-based broadcasting has the potential to better meet end-users sensory and cognitive media access needs than channel-based broadcasting as it can deliver separate audio objects.
- There are three main ways in which broadcast audio can be personalised to the benefit of hearing impaired populations: the speech to background ratio, the amount of redundant cues and the spatial separation between different elements of the audio.

- The majority of current work in object-based audio personalisation has explored speech to background ratio, with positive results. Balancing efficacy of this personalisation and ease of use by the end-user requires more effective methods for categorising and presenting audio objects.
- Personalisation of redundant non-speech audio objects shows potential but requires greater exploration of both the underlying psychoacoustic behaviours of normal and hard of hearing listeners and how interfaces for personalising this dimension could be designed.

Chapter 3

Characterising end-user needs

3.1 Introduction

Chapter 2 explored previous and ongoing research in broadcast accessibility. Chapter 3 aims to evaluate and define the current needs of end-users. The purpose of this is twofold; first to provide up to date understanding of the broadcast needs of both normal and hard of hearing viewers and, secondly, to guide the remainder of this work by identifying key areas where new audio technology could be best utilised to improve the end-user experience.

These aims are achieved through a survey to characterise end-user needs. This chapter describes the development, implementation, and results of this survey. The chapter begins by outlining the development of the survey instrument. This includes a brief review of the *Speech, Spatial and Qualities of Hearing Scale* [249] and its short form, the SSQ12 [250], which comprises part of the developed survey instrument (found in Appendix B). The next section describes the results from the survey which comprises of descriptive statistics of the quantitative data including respondent demographics, hearing abilities, and television watching habits, as well as directed content analysis of the qualitative data. Principal component analysis of the quantitative data is then performed. Finally, from these results and the results of Chapter 2, two research questions are developed which define the scope of this doctoral work.

3.2 Survey development

The aim of this chapter is to characterise the broadcast needs of a wide range of end-users. A survey methodology was selected as it allows for data acquisition from a broad range of individuals. As survey instruments are usually self-report, a large sample size is required in order to ensure the reliability of the data. To achieve a large sample size, online implementation of the survey was selected.

Numerous studies into broadcast needs have been conducted over the past two decades both in the UK and Europe [251, 100, 101, 15, 118, 122, 123, 21]. Whilst these studies establish a foundation for understanding end-user needs, the majority of these studies are more than five years old. As such, they will not account for changes and recent trends in programming and production style as well as advancements in broadcast technology. Furthermore, the previous work lacks the specificity on the hearing ability of respondents required to determine the relationship between different severities and types of hearing loss and subsequent broadcast needs. For this reason, the inclusion of an established and validated measure of hearing ability into the developed survey instrument was warranted and the Speech, Spatial and Qualities of Hearing Scale was selected [249, 250]. The remainder of this section describes the development of the survey instrument.

3.2.1 Demographics and hearing ability

The survey section 'About You' contains items to determine demographic, language, and self-identified hearing ability (which can be seen in Appendix B). These items generated supplementary variables to explain patterns within the response data (see Section 3.2.3 and 3.2.2). There are a number of studies debating whether musical training has an effect on an individuals' speech in noise performance [252–254]. To determine whether musicianship had any effect in the results here, respondents were also asked whether they were professional or amateur musicians.

To capture information about the self-identified hearing ability of respondents, they were asked to first indicate whether they identified as hard of hearing and, if so, encouraged to give further information about their hearing loss in the survey section 'About Your Hearing'. Here respondents could indicate the severity of their hearing loss, whether they have tinnitus and whether they use an assistive hearing device. If so, they could indicate the type, how regularly they utilise the device and how long they have had it fitted.

3.2.2 Speech, Spatial and Qualities of Hearing Scale

About the Instrument

The Speech, Spatial and Qualities of Hearing scale is a survey designed to capture information about an individual's perceptions of how their hearing (dis)ability impacts upon them in the complex listening situations encountered in everyday life [249]. It has been widely used as both a clinical assessment of the benefit of different hearing interventions (including, but not limited to, cochlear implants, bilateral hearing aids and bone-anchored hearing aids[255]) as well as in research investigations of younger and older listeners [256]. The survey can be self-administered or administered by a clinician or researcher.

The original, full form of the survey has 50 items although one question applies only to hearing aid users and is often omitted [256, 250]. The 49 item version is abbreviated here as SSQ49. SSQ49 consists of three categories of questions:

- Speech Hearing – 14 items
  - e.g. *"Can you have a conversation in the presence of someone whose voice is the same pitch as that of the person you're talking to?"*
- Spatial Hearing – 17 items
  - e.g. *"Do you have the impression of sounds being exactly where you would expect them to be?"*
- Qualities of Hearing – 18 items
  - e.g. *"Do you find it easy to recognize different people you know by the sound of each one's voice?"*

Each item asks the respondent to rate their (dis)ability to perform an auditory based 'activity' on an 11 point continuous scale from 0 (= complete disability) to 10 (= complete ability). The scores are usually reported in one of three ways: as individual item scores, as averages for each of the categories or averages across the whole instrument. An alternate approach has been to use ten 'pragmatic' subscales, developed by Gatehouse [257], which group the items based on a reasoned interpretation of the items rather than their statistical or psychometric properties. Factor analysis has also been utilised by Akeroyd to group the questions yielding three main factors closely relating to section of the SSQ49: 'Speech Understanding', 'Spatial Perception and Clarity' and 'Separation and Identification' [255]. Akeroyd also proposed a possible fourth factor: 'Effort and Concentration'.

One study has previously assessed the test-retest reliability of the SSQ49 finding that the test-retest correlation was 0.83 when the instrument was administered by a clinician in an interview setting and reduced to 0.65 when self-administered [258]. Internal reliability was

evaluated using Cronbach's alpha which is a measure of how closely related a set of items are to the underlying construct they are designed to measure. High values of Cronbach's alpha $\geq$0.9, indicate excellent internal consistency. Cronbach's alpha has been measured for the SSQ49 by two studies which showed that when administered by interview Cronbach's alpha was high (0.96-0.97 [258] and 0.96 [259]). The Cronbach's alpha was shown to be lower, 0.88-0.93 [258] when the instrument was self-administered though this still indicates good internal consistency (values $\geq$0.8 being defined as 'good').

The relationship between audiometric threshold hearing loss and results of the SSQ49 has been reported in work by Akeroyd [255]. Regression showed that, when data is grouped into unaided, unilaterally aided and bilaterally aided, the SSQ49 explains 85% of the variance. A much smaller amount of the variance, only 21%, is accounted for when analysis is performed on individual scores. Normative data for young, normal hearing respondents has also been gathered for the SSQ49 [260]. This indicated that normal hearing listeners will often not rate their listening abilities at the top of the scale and ratings substantially differ between respondents. Ratings from young, normal hearing populations show good agreement across studies [259, 261] and have mean ratings consistently higher than those obtained from hard of hearing populations [260].

There are also reduced versions of the SSQ49, ranging from only 5 items [259] to 12 [250] items. These reduced forms have many items in common with each other [255]. The authors of the SSQ5 demonstrated that even the use of only 5 items had greater sensitivity for identifying hearing loss (defined in their work as a speech reception threshold in noise of greater than -5.8dB SNR) than the question 'Do you have a hearing loss?'.

SSQ12

The SSQ12 was developed with the aim of being used as a short clinical tool, possibly in conjunction with other scales [250]. The items were selected based on whether they had: strong weightings on each of the three factors evident in Ackroyd's factor analysis of the SSQ49 [255], easy to visualise contexts, produced a wider range of scores and did not attract many 'not applicable' responses. The selected items can be seen in Table 3.1. Good agreement between the average results of the SSQ12 and original SSQ49 were found when implemented with a cohort of 1220 hard of hearing respondents (comprising 386 who were unaided, 627 who were unilaterally aided and 207 who were bilaterally aided). The SSQ12 however did have, on average, a modestly lower score. Whilst the authors in [250] do not give a figure for correlation between the scores, they model the relationship with the power function in Eqn. 3.1.

Table 3.1 Reduced 12 item form of the Speech, Spatial and Qualities of Hearing Scale and their corresponding 'pragmatic' subscales and major factors. Reproduced from Noble et al. [250].

ID	Question	Pragmatic Subscale [250]	Major Factor [249]
SSQ1	You are talking with one other person and there is a TV on in the same room. Without turning the TV down, can you follow what the person you're talking to says?	Speech in Noise	Speech Understanding
SSQ2	You are listening to someone talking to you, while at the same time trying to follow the news on TV. Can you follow what both people are saying?	Multiple Speech Streams	Speech Understanding
SSQ3	You are in conversation with one person in a room where there are many other people talking. Can you follow what the person you are talking to is saying?	Speech in Speech	Speech Understanding
SSQ4	You are in a group of about five people in a busy restaurant. You can see everyone else in the group. Can you follow the conversation?	Speech in Noise	Speech Understanding
SSQ5	You are with a group and the conversation switches from one person to another. Can you easily follow the conversation without missing the start of what each new speaker is saying?	Multiple Speech Streams	Speech Understanding
SSQ6	You are outside. A dog barks loudly. Can you tell immediately where it is, without having to look?	Localisation	Spatial Hearing
SSQ7	Can you tell how far away a bus or a truck is, from the sound?	Distance and Movement	Spatial Hearing
SSQ8	Can you tell from the sound whether a bus or truck is coming towards you or going away?	Distance and Movement	Spatial Hearing
SSQ9	When you hear more than one sound at a time, do you have the impression that it seems like a single jumbled sound?	Segregation	Qualities of Hearing
SSQ10	When you listen to music, can you make out which instruments are playing?	Identification of sound	Qualities of Hearing
SSQ11	Do everyday sounds that you can hear easily seem clear to you (not blurred)?	Quality and Naturalness	Qualities of Hearing
SSQ12	Do you have to concentrate very much when listening to someone or something?	Listening Effort	Qualities of Hearing

$$SSQ_{12} = 10\left(\frac{SSQ_{49}}{10}\right)^{1.25} \tag{3.1}$$

In order to make Eqn. 3.1 more readily interpretable, this relationship was modelled using randomly sampled values from a normal distribution with $\mu = 5$ and $\sigma = 1.5$. Linear regression gives a Pearson's R = 0.998 indicating high levels of agreement between the short and long forms of the instrument for this model of respondents.

Implementation

The SSQ12 was utilised as part of this instrument as it is a validated and reliable tool for assessing the effects of hearing loss. The shortened form was selected as it allowed for inclusion within the survey instrument without significantly increasing its length though still maintaining the majority of the advantages of the SSQ49. Furthermore, the sub-scales utilised by the shortened form represent the majority of types of listening scenarios to be expected within television content. This would facilitate comparison between an individual's perception of their ability to understand speech in everyday situations and their ability to understand speech on television. The original forms of the SSQ12 have a continuum of response values between 0 and 10. Due to the limitations of the online survey platform onlinesurveys.co.uk which was used, only discrete integer values were given as possible responses.

3.2.3 Experiences of television

To fully characterise the experiences of viewers and the problems they encounter with television speech, both closed- and open-ended qualitative questions were used as well as quantitative Likert scale questions (in the same form as the items in the SSQ12). These questions can be seen in Tables 3.2 and 3.3. Two of these items, both qualitative closed-ended question, were included in the survey section 'About You'. The remainder of these items, both quantitative and qualitative, made up the entirety of the survey section 'Your Experience of Television'.

The two closed-ended qualitative questions in the survey section 'About You' aimed to characterise a respondent's television viewing habits. This was in order to determine whether patterns in the data may be due to significant differences between the amount of television respondents watch per day or from watching vastly different types of programming. The selection of programming genres used were informed by the categorisation of programmes utilised by BBC iPlayer.

Table 3.2 Quantitative Likert scale items about television experiences developed for this work and their corresponding 'pragmatic' subscales and major factors. Responses are gained on a 11 point scale.

ID	Question	Pragmatic Subscale	Major Factor
TV1	Generally, how difficult do you find it to understand speech on television?	Television Speech	Speech Understanding
TV2	A character is speaking but they are not on screen. How easily can you understand the speech without seeing the character's face?	Lip-reading	Speech Understanding
TV3	You are watching a panel show and one of the panellists is speaking whilst the studio audience laughs and cheers. How easily are you able to understand the panellist's speech?	Speech in Noise	Speech Understanding
TV4	How often do you use subtitles?	Subtitle Use	N/A
TV5	A news presenter is reporting from a quiet studio. Without using subtitles, how easily can you understand the speech?	Speech in Quiet	Speech Understanding
TV6	You are watching a scene on television which has the sound of clinking glasses, music and people talking in the background. Can you make out the different sounds?	Identification of sound	Qualities of Hearing
TV7	You are watching a nature documentary. The narrator is speaking with the constant sound of a waterfall in the background. Can you follow what the narrator is saying?	Speech in Noise	Speech Understanding
TV8	How much effort do you require to hear what is being said in a television drama?	Listening Effort	Qualities of Hearing
TV9	How often do you watch television programs which are not in your native language?	Language: Non-Native Viewing	N/A
TV10	When sign-interpretation is available, how often do you watch sign language interpreted programming?	Language: BSL	N/A

Table 3.3 Closed- and Open-ended qualitative survey items addressing television experiences, including the possible responses for the closed-ended items.

Closed-ended qualitative questions
Q1 – How many hours a day do you watch television, on average? **Answers** – 'Less than 1 hr', '1 - 2hrs', '3 - 4 hrs', 'More than 4 hrs' **Q2** – What type of programming do you mostly watch? (you may select more than one option) **Answers** – 'News and Current Affairs programming', 'Drama and Soaps', 'Sports', 'Lifestyle, Music and Food', 'Documentary', 'Comedy', 'Films' **Q3** – Thinking about a recent drama you have watched on television. Which of the following sounds helped you personally to follow the plot? (Select as many as you think apply) **Answers** – 'Dialogue', 'Foreground sounds (e.g.the main character slamming a door in anger or other sounds that the characters can hear)', Background sounds (e.g. sounds of birds in the countryside, background chatter in a pub scene), Music **Q4** – Which of the following types of television content do you find the dialogue/speech easiest to understand whilst watching? (Select as many as you think apply) **Answers** – Same as Q2
Open-ended qualitative questions
Q5 – What measures do you feel would make television speech easier for you to understand? **Q6** – Are there any specific examples of times when you have found television speech hard to understand that you would like to tell us about?

Two more closed-ended qualitative questions were included in the survey section 'Your Experience of Television'. The first of these investigates the different broadcast elements which respondents feel help them to follow the plot of a programme by asking respondents to recall a recent drama they have watched. A previous study has demonstrated that for some hard of hearing listeners, foreground sounds aid their understanding of the narrative and subsequently support speech intelligibility [170]. The categories of broadcast elements used are based on those utilised in the study by Shirley [170] based on work categorising broadcast objects by Woodcock [227]. Drama was selected as it is a genre in which all these objects types are regularly present. The second closed-ended item aimed to to determine which genres respondents find speech easiest to understand. Along with the programming types most watched by respondents, this allowed determination of whether there was some content which, due to style and production, is easier to understand or whether genres which people more commonly watch were easier to understand (due to familiarity and priming). This item used the same genres which were used in the item about most watched types of programming.

Two open-ended qualitative items were included to allow individuals to explore ideas beyond the structured scope of the survey and relate any other problems or experiences that they may wish to express. One of these questions focused on previous experiences of problems with speech intelligibility on television whilst the other requested respondents to envisage what measures would improve their experience of television sound.

The quantitative items in this section were designed to cover four of the pragmatic subscales utilised by the SSQ12, two from the major factors 'Speech Understanding' and two from 'Speech Qualities'. To facilitate comparison, they were phrased in the same manner as the SSQ12 items and used the same 11 point scale. This allowed any differences between the scores to be attributed to the content of the question rather than its phrasing. They can be seen in Table 3.2. One of the subscales not utilised in the SSQ12, 'Speech in Quiet', is also added. This is as lacking this subscale was identified by the developers of the SSQ12 as one of the key potential factors in the differing average scores between the SSQ12 and the SSQ49. Additionally, it added four television based subscales: 'Television Speech', 'Language', 'Subtitle Use', and 'Lip-reading'. A general item on television speech, TV1, was also included.

3.2.4 Additional items

A number of additional items were included in the survey instrument to help identify problematic data, problems with the presentation of the online survey on some devices, data anonymity, and to offer respondents the opportunity to sign up for further information

and research participation. These included the survey sections 'Taking Part' and 'Further Participation' as well as question 8 in the 'About You' (seen in Appendix B).

3.2.5 Participation

The survey was completed online by 125 people over a period of eight months (June 2017 to January 2018). One participant was omitted due to errors in their submission. Throughout these results, any quantitative responses which were omitted by the respondent (constituting 0.2% of the total quantitative responses) have been imputed to the mean of that item.

Participants were self-selecting and were recruited through a wide range of methods in order to gain a varied participant pool. Advertisements to the general public were made through University media channels and BBC Radio Manchester. Participants were also recruited through professional organisations like the Institute of Acoustics and charities like the National Association for Deafened People.

3.3 Results: Demographics and hearing ability

3.3.1 Age and hearing ability

Fifty-five respondents were over 50 years of age, one declined to give their age and the remaining 68 participants indicated an age between 18 and 50 years.

Ninety-one respondents identified as normal hearing, 32 identified as hard of hearing and one declined to identify their hearing ability; Table 3.4 shows the severity of hearing loss indicated. The majority had moderate hearing loss and exactly half of hard of hearing respondents, 16 respondents, also had tinnitus. Twenty of the hard of hearing respondents indicated that they used assistive hearing devices, half were bilateral aids and half unilateral. These devices were predominantly hearing aids of varying types, worn by 18 respondents, whilst one respondent had a cochlear implant and one had a bone anchored hearing aid. The majority of these, 14 respondents, use their device regularly. Respondents who used assistive hearing devices had had them fitted for a mean duration of 16.6 years with the most recent being fitted 4 months ago and the longest being first fitted over 50 years ago.

The mean value of the SSQ12 questions for hard of hearing listeners was 3.8 ($s = 2.9$) whilst for normal hearing listeners it was 7.3 ($s = 2.3$). These are significantly different at the level $[p < 0.001]$, utilising a two tailed Wilcoxon Rank Sum Test. These averages are similar to the normative data from the full SSQ49 [260] indicating that the quantisation of the results scale likely had limited impact. The relationship between the SSQ12 average and self-reported hearing severity was also assessed, using a one way ANOVA, indicating

Table 3.4 Self-identified severity of hearing loss of participants who identified as hard of hearing, noting the number of participants in each severity level and the percentage of total hard of hearing respondents that this represents.

Severity	Count	Percentage
Mild	7	21.88%
Moderate	12	37.50%
Severe	9	28.13%
Profound	3	9.38%
Not Sure	1	3.13%

that hearing ability was a statistically significant factor in determining respondent's average SSQ12 score $[F = 6.06, df = 1, p = 0.003]$.

3.3.2 Country of residence and native language

Respondents were from seven different countries, with 100 respondents from the United Kingdom (82.6%), 14 respondents from Australia, two from the United States, two from Spain and one from each Switzerland, Canada and Denmark. One respondent declined to identify their country of residence. The majority, 114 respondents, identified their native language (first language spoken) as a variant of English. Other native languages identified were Catalan spoken by two respondents as well as Czech, Tamil, Portuguese, and Afrikaans which were spoken by one respondent each. Four respondents did not identify their native language.

3.3.3 Musicianship

All respondents answered this question with just over half, 69 respondents, identifying as not being musicians. Of those remaining, 51 respondents identified as being amateur musicians and four identified as being professional musicians.

3.3.4 Discussion

The majority of respondents identified as being resident in the UK and subsequently the representativeness of the data will be considered with reference to the UK population. In 2015, approximately 16% of the UK population had some degree of hearing loss [3] and 28% of respondents identified as having some hearing loss which is above the average in the UK population. Of those, 60.7% identified as having mild to moderate loss. This is a

much smaller proportion of severely hearing impaired individuals than the general population where mild-moderate loss makes up 91.7% of all those with hearing loss [3]. A further 32.1% identifying as having severe hearing loss and 7.1% identified as having profound hearing loss. Of those respondents who had an assistive device fitted, 66.6% identified that they utilised it regularly, which is higher than the average in the general hearing impaired population [28]. However, it is likely that this is significantly influenced by the higher incidence of loss in the respondents as individuals with greater hearing loss are more likely to depend on assistive devices. The average age of participants was 47.0 which is skewed younger than the average age of the general population (due to an ageing population). These differences between the general population and the sample population were considered when interpreting the results. Figures for the percentage of the population who are musicians could not be identified; however as the sample is almost evenly balanced between musicians and non-musicians, this was not considered to be an issue.

There were 17.4% of respondents who were not resident in the UK and, given the limited size of this group, their representation of their respective countries could not be determined. In order to determine whether their countries of residence resulted in any significant difference in their response data, investigations were undertaken using principal component analysis (PCA) in Section 3.5.

3.4 Results: Experiences of television

3.4.1 Television viewing habits

Average hours of television watched per day was grouped by age and expressed as a percentage of the respondents in that age group. In all age groups, the majority watched 1–2 hrs per day. The proportion of each age group in the two higher categories (3–4 hrs, More than 4 hrs) decreased monotonically from the highest proportion in the 65+ age group (35%) to the lowest in the youngest age group (22%). This is consistent with audience research that shows older viewers watch more hours of television [8].

Figure 3.1 shows the most commonly watched types of programming identified by respondents. Over 50% of respondents identified ‘Films’, ‘Comedy’, ‘Documentary’ and ‘News and Current Affairs programming’ as contributing to their most commonly watched programming. ‘News and Current Affairs programming’ was the most regularly watched genre with over 60% of respondents selecting it.

3.4.2 Closed-ended qualitative data

Figure 3.1 also shows the programme genres in which speech was identified as the easiest to understand and 'News and Current Affairs programming' was identified by over 80% of respondents. This was followed by 'Comedy', with over 50% of respondents. 'Films' (8.8%) as well as 'Drama and Soaps' (8.0%) were identified by the lowest proportions of respondents as having easy to understand dialogue.

When asked to consider a recent drama they had watched and to identify the types of sounds which helped them follow the plot, the majority of normal and hard of hearing respondents identified 'Dialogue' (96.7% and 81.3% respectively). A greater percentage of normal hearing respondents identified 'Foreground Sounds' (NH: 49.5%, HoH:25.0%), 'Background Sounds' (NH:24.2%, HoH:15.6%) and 'Music' (NH: 40.7%, HoH:12.5%) as helping them to follow the plot than did hard of hearing respondents. For both groups, more respondents identified 'Foreground Sounds' as helping them to follow the plot than other non-speech sounds.

3.4.3 Quantitative data

Sixteen respondents identified utilising sign-interpretation when it was available but only one indicated that they regularly watched it (rated greater than 5 on the Likert scale). Thirty-four respondents identified that they sometimes watch television programmes which are not in their native language. Of these 34 respondents, 28 had English as their native language. Those who regularly watch programmes not in their native language (rated greater than 5 on the Likert scale) included 11 respondents with English as their first language and 5 respondents with other native languages.

The mean results for the remaining eight quantitative items about experiences of TV are summarised in Table 3.5. Two questions from the SSQ12 which pertained to listening in scenarios where television speech is present are also reported here. These were computed separately for normal hearing and hard of hearing respondents. The differences between the groups were assessed using the Wilcoxon Rank Sum Test.

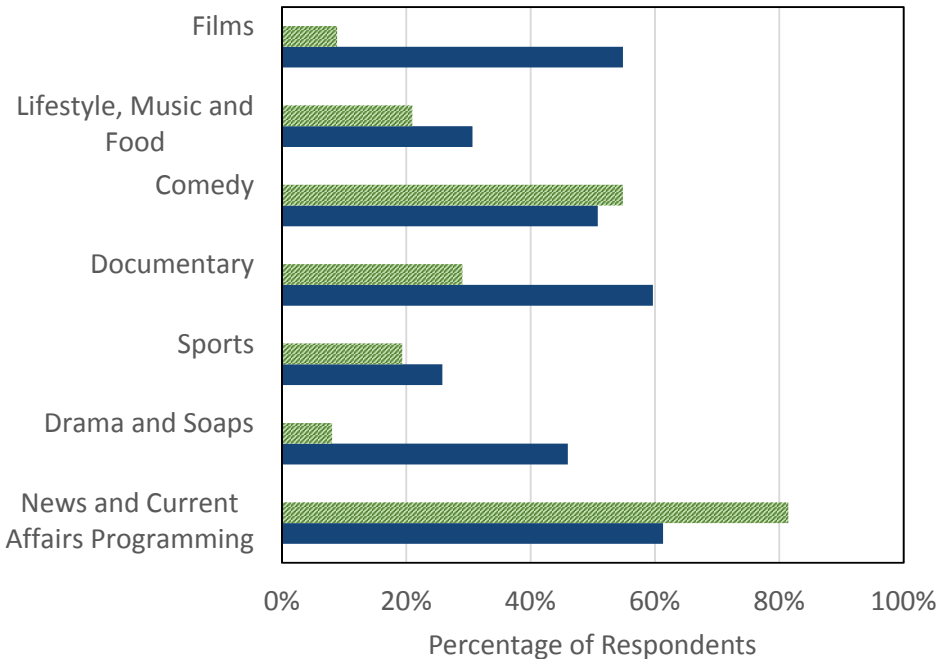
As can be seen from Table 3.5, hard of hearing listeners scored significantly lower than their normal hearing counterparts for all items except subtitle usage (TV4)[$p < 0.001$]. Whilst a significant difference exists between subtitle usage, it is at a lower significance level [$p < 0.05$]. This was further investigated and showed that 62.0% of normal hearing respondents sometimes used subtitles (rated non-zero on Likert scale). Of these, 79.9% identified watching some foreign language content leaving a group of respondents who identify as normal hearing and do not watch foreign language content but who utilise

Table 3.5 Mean values of the quantitative items from survey section *'Your Experience of Television'* and the first two items of the SSQ12, shown for hard of hearing and normal hearing respondents respectively (standard deviation noted in parentheses).

ID	Survey Items	Hard of hearing	Normal Hearing
Television Experience			
TV1	Generally, how difficult do you find it to understand speech on television?	3.6 (2.4) ***	7.6 (2.1) ***
TV2	A character is speaking but they are not on screen. How easily can you understand the speech without seeing the character's face?	3.3 (2.8) ***	7.2 (2.0) ***
TV3	You are watching a panel show and one of the panellists is speaking whilst the studio audience laughs and cheers. How easily are you able to understand the panellist's speech?	2.7 (2.3) ***	6.2 (1.8) ***
TV4	How often do you use subtitles?	4.7 (4.2) *	2.7 (3.1) *
TV5	A news presenter is reporting from a quiet studio. Without using subtitles, how easily can you understand the speech?	6.7 (3.1) ***	9.1 (1.5) ***
TV6	You are watching a scene on television which has the sound of clinking glasses, music and people talking in the background. Can you make out the different sounds?	4.6 (2.9) ***	7.6 (1.8) ***
TV7	You are watching a nature documentary. The narrator is speaking with the constant sound of a waterfall in the background. Can you follow what the narrator is saying?	3.9 (2.9) ***	7.7 (1.8) ***
TV8	How much effort do you require to hear what is being said in a television drama?	2.4 (2.4) ***	6.8 (2.3) ***
SSQ12			
SSQ1	You are talking with one other person and there is a TV on in the same room. Without turning the TV down, can you follow what the person you're talking to says?	3.3 (2.5) ***	7.3 (2.1) ***
SSQ2	You are listening to someone talking to you, while at the same time trying to follow the news on TV. Can you follow what both people are saying?	2.2 (2.0) ***	5.2 (2.4) ***

$^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$

Figure 3.1 Responses to the questions: *'What type of programming do you mostly watch?'* (noted by blue bars) and *'Which of the following types of television content do you find the dialogue/speech easiest to understand whilst watching?'* (noted by green bars) expressed as a percentage of total respondents.



subtitles. Furthermore, of those who identified as hard of hearing, only 68.7% used subtitles. Four of these items refer to speech in a particular genre of programming. The higher than average responses to TV5 indicate that speech in news is easier to understand which is reflected in Figure 3.1. The genre which scored the lowest for both groups on average was comedy panel show even lower than documentary. This differs from Figure 3.1 where more people identified 'Comedy' than 'Documentary' as having easily understood speech.

3.4.4 Open-ended qualitative items

Qualitative content analysis

Content analysis was selected to analyse the results from the two open-ended qualitative items. Qualitative content analysis is a research methodology for describing and interpreting textual data through a systematic process of coding [262]. Content analysis has been used widely for analysing qualitative data, including interviews and observational data, in the field of health for the past half century. The term content analysis can be used to refer to a variety of methods which range from impressionistic and intuitive through to systematic and strict textual analysis [263]. The term as used here refers to the latter. Any kind of content analysis

Table 3.6 Pairwise Spearman Rank Correlation Coefficient for each pair of raters for each question from Table 3.3.

	Q5 (from Table 3.3)		Q6 (from Table 3.3)	
	1	2	1	2
2	0.8314 ***	–	0.7346 ***	–
3	0.8386 ***	0.7940 ***	0.7799 ***	0.5997 ***

$^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$

has limitations as it relies on human coders who are inherently subjective. However, this is mitigated by the increased value obtained through human coders as opposed to objective techniques like text mining. Using human coders yields more meaningful data through providing an analysis of not only the words themselves but the context of the response as a whole as well as the question.

The type of content analysis selected for analysing the free text data was directed content analysis using the unconstrained matrix approach [264, 262]. Directed, or deductive, content analysis utilises a framework of established theory or concepts on the research topic into which each of the responses is allocated. This differs from inductive content analysis where the matrix of codes is developed by the coders during the process of coding the responses [262] or Grounded Theory which is another inductive approach to qualitative data analysis [265]. Inductive approaches are most appropriate for research areas with little established theory or where there is no hypothesis being tested. Given the established theory in broadcast speech intelligibility [21, 104], and the additional robustness of deductive analysis to variation in individual coders, this approach was selected.

The main issue deductive analysis presents is that it does not accommodate for new developments in theory. An unconstrained matrix approach rectifies this by primarily basing the analysis on existing theory but allowing for the addition of new categories if warranted [264].

Analysis approach

The coding framework was developed from an independent source, Armstrong's *'Audibility Problem Space'* [21] which is reproduced in Table 3.7. This particular framework was selected as most suitable as it is specific to English language broadcast content, is recent (published in 2016) and is slightly more detailed than the problem space defined by Mapp

Table 3.7 Coding framework used for directed content analysis of qualitative items, based on the *'Audibility Problem Space'* (reproduced from Armstrong [21] with the additions from Section 3.4.4 shown in red text).

Audibility Problem Space				
1. Production /Direction	**2. Capture**	**3. Post Production**	**4. Broadcast**	**5. Home**
.1 Writing style – complexity of narrative .2 Direction and camerawork .3 Choice of shots – showing actor's face at vital moments in the narrative .4 Voices, accents, dialects & clarity of delivery .5 Choice of location - control of background noise .6 Multiple speakers at a time	.1 Choice of microphone .2 Microphone positioning .3 Skill of sound recordist .4 Manual level control vs agc/limiter .5 Priority given to sound over video – e.g. microphone allowed in shot .6 Retakes for sound .7 Voiceover booth matching to field recordings - ADR .8 Use of digital compression and codec choice (mobile phone audio!)	.1 Voice processing, equalisation, noise reduction /gating, level control/limiting .2 Added music, level, type, purpose, style .3 Added sound effects, level, type purpose .4 Added voiceover .5 Video edit, choice of shots .6 Loudness control .7 Dynamic range .8 Stereo or 5.1 mix .9 Inclusion of open subtitles .10 Balance between audio elements	.1 Output processing or other dynamic range control - not currently used on BBC TV .2 Audio encoding quality and any cascaded coding .3 Quality and availability of closed subtitles	.1 Television sound quality – receiver, amplifier and speakers .2 Use of mono, stereo or 5.1 loudspeakers .3 Room Acoustics and the positions of the television and the viewer in the room .4 Background noise .5 Viewer's hearing - level of interest in the programme - knowledge of the programme's subject matter - expectations of the programme - language skills - willingness to use subtitles .6 User control of reproduction

[104]. The items in the problem space were given identifying codes, based on their categories, to simplify the process of coding.

For responses which could not effectively be reconciled to a single (or small number of) codes, a new code was generated, extending the problem space. These extensions are shown in Table 3.7 in red text. Three new codes were generated: 1.6 – MULTIPLE SPEAKERS AT A TIME, 3.10 – BALANCE BETWEEN AUDIO ELEMENTS and 5.6 – USER CONTROL OF REPRODUCTION. 3.10 – BALANCE BETWEEN AUDIO ELEMENTS was created to reconcile the multiple codings of responses like *'Quieter background sounds in relation to speech'*. 1.6 – MULTIPLE SPEAKERS AT A TIME… was created to encompass both style and genre, e.g. "Question Time" which, due to the nature of politics, regularly includes individuals speaking over each other, as well as performance facets of this problem. 5.6 – USER CONTROL OF REPRODUCTION was created to address problems identified as being caused by insufficient

ability for end-users to control how content is reproduced in their home, such as *'Having the option to switch off background music or sounds that are not relevant to the topic or action'*. Additionally, a modification was made to code 4.3 – QUALITY OF CLOSED SUBTITLES to encompass availability.

For responses which were non-specific, for example *'Increased Volume'*, it was assumed, based on the phrasing of the questions, that these comments referred to speech. Responses which did not identify any tangible problems (e.g. the response *'Happens regularly'*) were omitted. It also should be noted that many of the codes do not specifically indicate the magnitude or direction of the problem. Items coded as 3.7 – DYNAMIC RANGE can equally refer to responses such as *'less dynamic range'* as *'[...] the dynamic range is too great'*. The results given in the following section indicate only how many people believed that there is a problem with that facet of television capture, production or reproduction and should be interpreted accordingly.

Inter-rater agreement

After each rater had completed coding the responses, the agreement between the different raters was calculated by performing pairwise comparison using Spearman's rank correlation coefficient. The results of these for each pair of raters for the responses from each question can be seen in Table 3.6. Very good agreement can be seen between all coders for the first qualitative question and good agreement can be seen for the second.

Results of directed content analysis

There were 102 participants who responded to the question *'What measures do you feel would make television speech easier to understand?'* with an average of 1.85 codes assigned to each response (per rater). The null hypothesis for this analysis was that the 'responses are uniformly distributed' (i.e. all categories were selected an equal number of times in the responses). This hypothesis is representative of a case where all parts of the problem space (i.e. broadcast chain) are equally likely to be identified as needing improvement. Disproving the null hypothesis allows for areas to be identified which end-users perceived as having greater capacity to improve their experience.

To test this hypothesis, a Chi Squared Goodness of Fit test was used. The codes which attracted significantly more responses than others can be seen in Figure 3.2. The Chi Squared statistic is given by the Eqn 3.2, where k is the number of categories, n is the total number of responses and n_j is the number of responses in the j^{th} category. This was completed both for each the individual raters and the aggregate of all raters to explore whether some codes

were used by some raters more than others. First, a test statistic was calculated to determine whether the null hypothesis could be rejected.

$$z = \sum_{j=1}^{k} \frac{(n_j - \frac{n}{k})^2}{\frac{n}{k}} \tag{3.2}$$

The test statistic z follows a χ^2 distribution with $(k-1)$ degrees of freedom, therefore the critical interval for the test statistic is given by Eqn 3.3.

$$z \pm \chi^2(k-1)\frac{\alpha}{2} \tag{3.3}$$

The null hypothesis was rejected in all scenarios $[z \geq 315, p < 0.01]$ indicating code use was not distributed uniformly at chance frequency. To determine which individual codes were used at greater than and less than chance frequency, the standardised residual in Eqn. 3.4 was used, as in [266].

$$R_{\text{si}} = \frac{x_i - \frac{n}{k}}{\sqrt{\frac{n}{k}}} \tag{3.4}$$

Rearranging Eqn. 3.4 in Eqn. 3.5 gives the frequency of code use required to indicate a significant difference (C_{sig}), at the level $\alpha = 0.05$ giving a range of ± 1.96 for a two tailed test.

$$C_{\text{sig}} = \left(\pm 1.96 \cdot \sqrt{\frac{n}{k}} \right) + \frac{n}{k} \tag{3.5}$$

Codes which were used significantly more are shown in Figure 3.2 with 10 codes identified. The percentage allocation of that code (aggregated over the three raters) is shown. All three raters had the same three most significant codes:

- 1.4 – VOICES, ACCENTS, DIALECTS & CLARITY...,
- 3.10 – BALANCE BETWEEN AUDIO ELEMENTS
- 3.2 – ADDED MUSIC LEVEL....

Four codes were only significant for two out of three raters:

- 5.1 – TELEVISION SOUND QUALITY...,
- 1.5 – CHOICE OF LOCATION - CONTROL OF BACKGROUND NOISE,
- 4.3 – QUALITY & AVAILABILITY OF CLOSED SUBTITLE
- 3.3 – ADDED SOUND EFFECTS....

Only one rater had allocated 3.1 – VOICE PROCESSING, EQUALISATION... significantly more than other codes, when analysed individually.

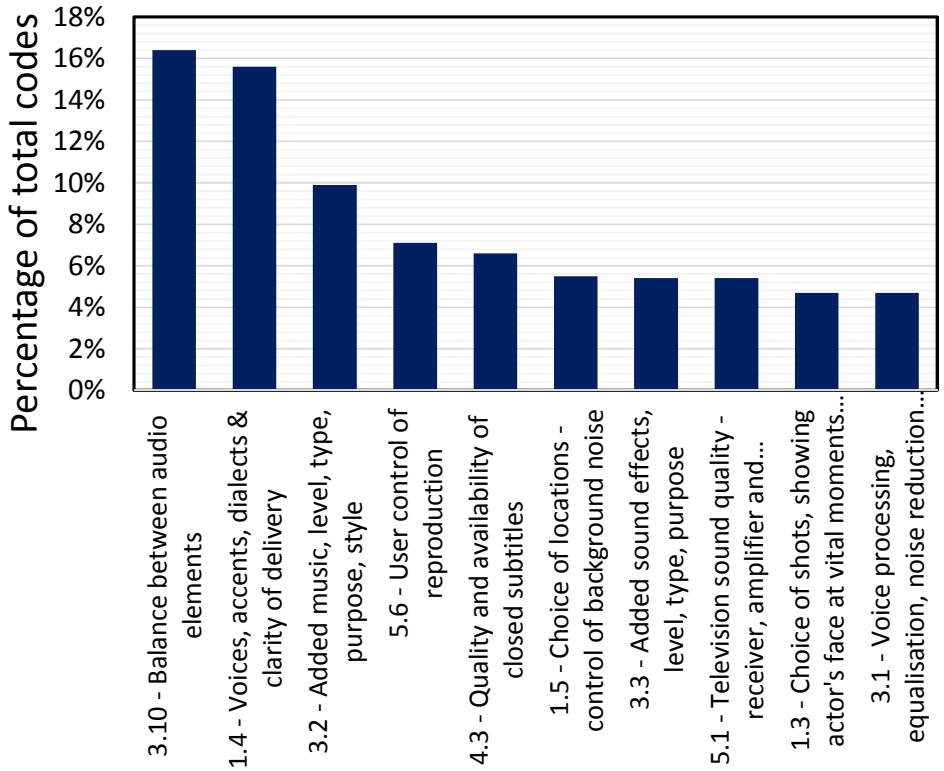


Figure 3.2 Codes allocated significantly more than chance [$p < 0.01$] in responses to the question Q5 (from Table 3.3: *'What measures do you feel would make television speech easier for you to understand?'* aggregated from all raters, shown from highest percentage allocation to lowest.

Eighty-two respondents answered to the question, *'Are there any specific examples of times you have found television speech hard to understand that you would like to tell us about?'*. Initial analysis of the responses indicated two main types of response; examples of problematic genres or specific programmes, and examples of the problems themselves. Some responses included both types. For the examples of the problems, they were treated in the same manner as for the previous open-ended question and coded using Table 3.7. This yielded an average 1.6 codes per response. A Chi Squared Goodness of Fit test was performed giving a test statistic of $z \geq 399$ $[p < 0.01]$, again rejecting the null hypothesis in all scenarios.

The significant codes from the second qualitative question can be seen in Figure 3.3. Given the lower inter-rater agreement for this question and the fewer number of responses, these results should be interpreted with greater caution. Three codes were identified as significant for all raters, with two of these being the most allocated codes for all raters. Only two raters had 4.3 – QUALITY AND AVAILABILITY OF CLOSED SUBTITLES and 1.6 – MULTIPLE SPEAKERS AT A TIME... identified as significant and three codes (5.5 – VIEWER'S HEARING..., 3.3 – ADDED SOUND EFFECTS... and 3.2 – ADDED MUSIC...) were only identified as significant for one rater, with 5.5 – VIEWER'S HEARING... and 3.3 – ADDED SOUND EFFECTS... not being significant overall. This is reflective of the lower inter-rater agreement.

To better capture the genres and programmes identified in the second qualitative question, a word cloud of the responses was generated using the `wordcloud` package in `R` [267]. The text data was first manually cleaned, which included rectifying obvious spelling errors (e.g. "Pole Dark" was rectified to the correct programme title, "Poldark") and concatenating programme titles and actor/presenters names to ensure they would be treated as a single concept. Cleaning was also performed using the `tm` package in `R` [268] to remove punctuation and stop words and to stem the text to their base word stems. This results in 549 words with a mean usage frequency of 3.2.

The results of this can be seen plotted in Figure 3.4 where all words with a frequency greater than one are shown. To aid readability and interpretability of the cloud, the word stems were plotted in their singular noun or verb form. Font size and colour co-very to show frequency, with high frequency words having large text size and low frequency words using a smaller size. The twelve most common words (in descending order, frequency in parentheses) were: drama (26), background (16), sound (16), dialogue (15), difficult (13), programme (13), lot (12), speech (12), accent (12), mumble (11), subtitle (11) and television (11).

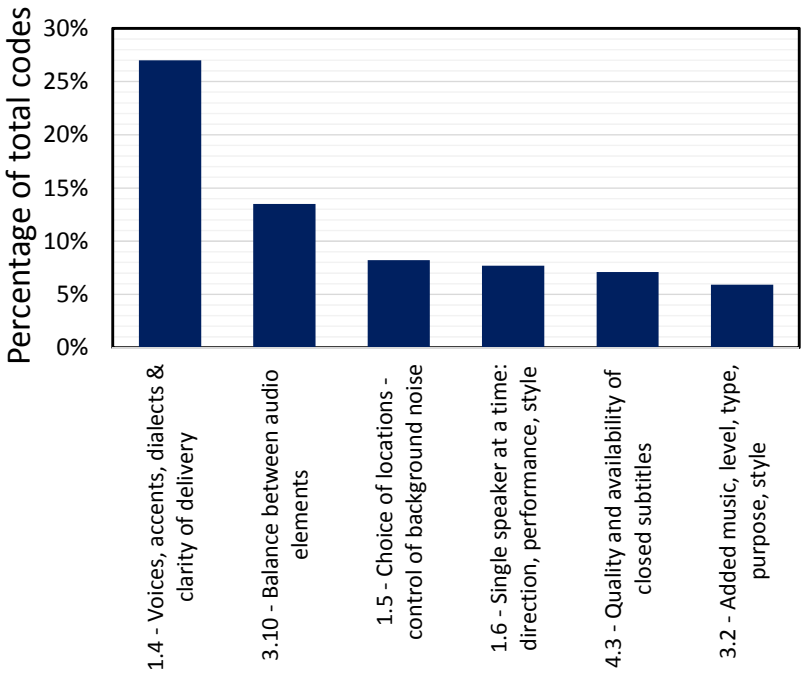
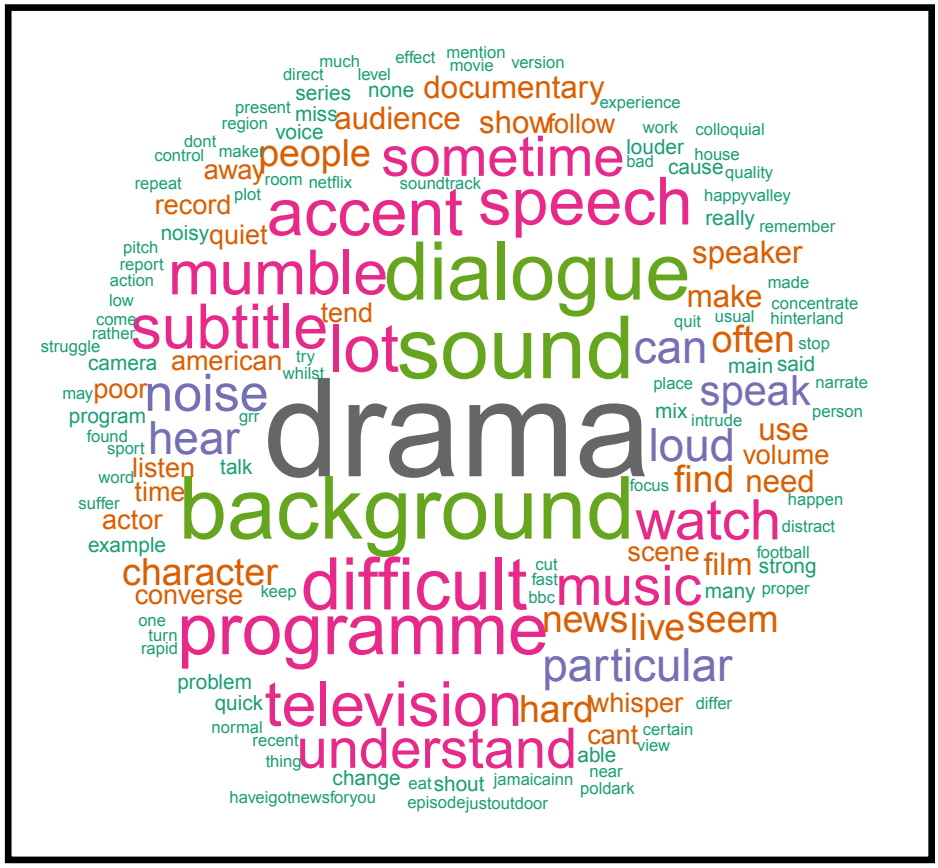


Figure 3.3 Codes allocated significantly more than chance in responses to the question Q6 from Table 3.3: *'Are there any specific examples of times you have found television speech hard to understand that you would like to tell us about?'* aggregated from all raters, shown from highest percentage allocation to lowest.

Figure 3.4 Word cloud for responses to the question Q6 in Table 3.3: *'Are there any specific examples of times you have found television speech hard to understand that you would like to tell us about?'*. Frequency is indicated by font size and colour.



3.4.5 Discussion

The quantitative results in Table 3.5 indicate that hard of hearing viewers find television speech significantly more difficult to understand overall than normal hearing viewers. This is an expected result. It is notable that there was only a weakly significant difference between subtitle usage showing that the rate of subtitle usage in the normal hearing population is only marginally less than for the hard of hearing population. This reflects previously reported increases in use of subtitles across age groups [188]. The low rates of subtitle use for hard of hearing listeners suggests that either an insufficient amount of content is subtitled or, alternatively, the available subtitles are not of sufficiently high quality. This argument is supported by the free text responses with 4.3 – QUALITY AND AVAILABILITY OF CLOSED SUBTITLES appearing significantly more than chance frequency in the responses to both questions. Furthermore, only a very small percentage of hard of hearing respondents identified using sign interpretation more than occasionally, indicating that it is unlikely that low subtitle use relates to increased use of sign-interpretation. The other results from the quantitative items are discussed within Section 3.5.4.

The open-ended qualitative results in Figure 3.2 and 3.3 verify and build on the results of previous investigations [101, 15, 104]. Five of the significant codes from Figure 3.2 correspond directly to the problems identified in the BBC study: 3.10 – BALANCE BETWEEN AUDIO ELEMENTS, 1.4 – VOICES, ACCENTS..., 1.5 – CHOICE OF LOCATION, CONTROL OF BACKGROUND NOISE, 3.2 – ADDED MUSIC... and 3.3 – ADDED SOUND EFFECTS.... The latter also directly reflects the third most common problem identified in the closed-ended questions in Strelcyk's work. 4.3 – QUALITY AND AVAILABILITY OF CLOSED SUBTITLES, significant in both Figures 3.2 and 3.3, reflects the free text responses in Strelcyk's study where it was the second most commonly identified issue. The identification of 5.6 – USER CONTROL OF REPRODUCTION goes beyond Strelcyk's findings; that those with greater overall problems have greater remote control usage, to demonstrate that there is a desire by the user to control and address problems themselves. 5.1 – TELEVISION SOUND QUALITY... complements the results shown by Mapp highlighting that variable quality and accuracy of television sound reproduction is having a significant negative effect on audiences. Additional problems are highlighted here which were not captured by previous studies: 1.3 – CHOICE OF SHOTS, SHOWING ACTOR'S FACE AT VITAL MOMENTS, 3.1 – VOICE PROCESSING... and 1.6 – MULTIPLE SPEAKERS AT A TIME....

Accent, delivery, and background noise (due to the location of recording) are facets of capture, and to an extent, production. As such, issues with these facets need to be controlled and accounted for during the production process, not afterwards. Some accessibility strategies,

like an increased speech to noise ratio, could mitigate these issues but the aim should be to remove the cause of these issues not compensate for them.

Television quality, in terms of loudspeakers, is a facet that can only be improved though greater focus on sound in the design process for televisions or through use of external, higher quality, sound systems in the home. The remaining facets are areas which object-based broadcast has significant potential to improve upon. In particular, object-based broadcasting methods can give the end-user control of how different objects are reproduced in the listening environment and allow them to adjust the balance of these objects based on their needs.

It can also be seen from Figure 3.4 that the genre 'Drama' dominates the problematic experiences with television speech. Film and documentaries are also mentioned with high frequency, which is consistent with the results in Figure 3.1. One possible reason that drama and film content is perceived as having the most difficult to understand speech is that it has the most complex soundscapes. This is supported by the fact that the easiest to understand types of content, 'News and Current Affairs programming', consists primarily of dialogue with comparatively minimal added background sounds or music. From this it appears that accessibility strategies for drama and film genres have the greatest potential for improvement and impact on end-user experience. It was also noted that "Netflix" appeared in Figure 3.4, indicating that speech intelligibility problems are not limited to terrestrial television media.

Aside from 'Dialogue', 'Foreground Sounds' were identified by both groups as being most useful to understanding the plot. This is consistent with results from previous studies [170]. This suggests that if speech is to be made clearer but the narrative is still to be effectively conveyed, these foreground sounds need to be given priority over other audio elements. This is corroborated by responses to the open-ended items such as *'Having the option to switch off background music or sounds that are not relevant to the topic or action'* which indicate that end-users not only want greater control over non-speech sounds, but wish to control groups of sound differently. 'Dialogue' was not identified by 2.7% of normal hearing and 18.7% of hard of hearing respondents as being vital to following the plot. This unusual result may be due to subtitle use for the hard of hearing respondents or may be due to the phrasing of the question.

3.5 Principal component analysis of results

To reduce the dimensionality of the data and investigate the relationship between respondents' experience of television viewing and experience of hearing in everyday situations, the quantitative data was subjected to principal component analysis (PCA). PCA is an approach used to analyse data sets which have a wide array of variables which may or may not be

correlated. From the data it generates a series of uncorrelated 'Principal Components' (PC) which the original data can be projected onto. This transformation process can be considered akin to transforming a sphere from Cartesian coordinates to spherical coordinates; what the coordinates represent, a sphere, is unchanged but the coordinates have been translated to a representation which is more meaningful, given the dataset.

By representing the data in a more meaningful way, patterns in the data can be identified. If the data is highly correlated, the representation can be reduced to a small number of principal components for interpretation. Each of these principal components is representative of a dimension of variance in the data and by probing which variables most significantly correlate with this dimension, or *load* onto the dimension, the source of this variance can be surmised[1].

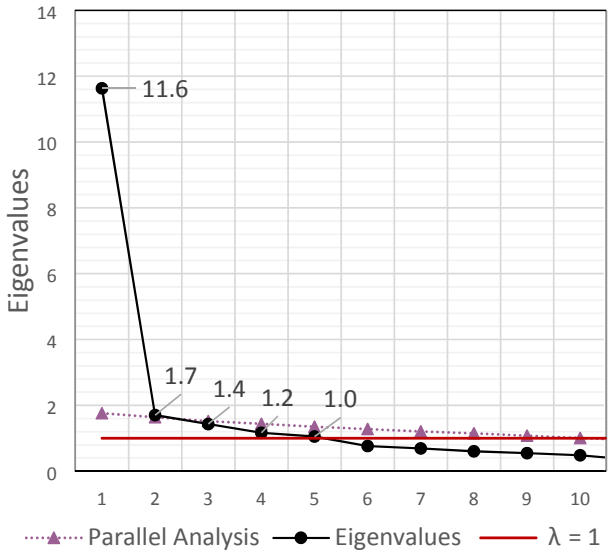
Given it is likely that the results from the SSQ12 and television questions are correlated, PCA is an appropriate method for determining the key sources of variance in the data. PCA selected, in preference to factor analysis, as it makes fewer assumptions about the underlying factor structure of the data and is a more psychometrically sound procedure [270]. Initial PCA showed that the majority of the variance was contained within the first principal component (explaining 50.6% of variance); subsequent components each explain less than 10% of variance. The eigenvalues from this initial analysis can be seen in Figure 3.5.

3.5.1 Significant dimensions

There are a number of methods which can be used to determine the significant dimensions of the PCA: eigenvalues (λ) greater than one (Kaiser's rule [271]), point of inflexion on the scree plot (determined visually or by an acceleration factor) and parallel analysis, which utilises random data in order to estimate which λ could have occurred by chance, suggesting that only λ above that threshold are significant [272, 255, 270]. Often multiple approaches are trialled and a comparison is made between them. Given the heuristic nature of the methods, however, parallel analysis has been shown to outperform many other methods [273]. All three methods were applied here and the results can be seen in Figure 3.5. It can be seen that the inflexion point method showed one significant dimension, parallel analysis showed two and Kaiser's rule indicated four. Given this disparity in the results, further investigations were made before selecting which dimensions to retain.

[1] For those unacquainted with PCA, the author recommends Shlens highly readable tutorial as a introduction to the method [269]

Figure 3.5 Eigenvalues for the first 10 dimensions, with eigenvalues above one noted (showing four significant dimensions) and eigenvalues from parallel noted in purple (showing two significant dimensions). The inflexion point of the scree plot can also be seen (showing one significant dimension).



3.5.2 Rotation and factor loadings

As in previous factor analysis of the SSQ items [255], rotation is used here to help elucidate the underlying factor structure of the data. Orthogonal rotation using varimax in R is used to rotate the factors. Akeroyd utilised oblique rotation, arguing it is highly unlikely that the different components of hearing ability are uncorrelated [255]. Both types of rotation were trialled in this work and it was seen that, for this data, oblique (using promax in R from the psych package [274]) and orthogonal rotation produced similar results. As orthogonal rotation yields more interpretable explained variance, given the dimensions are not correlated, this approach was selected.

Table 3.8 Percentage of total variance in the data contained in the first unrotated principal component and 2 – 4 components, after orthogonal rotation. Note, that the addition of the fourth dimension offers only a small portion of explained variance in the fourth dimension.

# of PCs	Dim. 1	Dim. 2	Dim. 3	Dim. 4	Total Variance
1	50.6%	–	–	–	50.6%
2	30.1%	27.8%	–	–	58.0%
3	10.4%	28.6%	25.2%	–	64.2%
4	23.6%	20.3%	19.4%	5.9%	69.3%

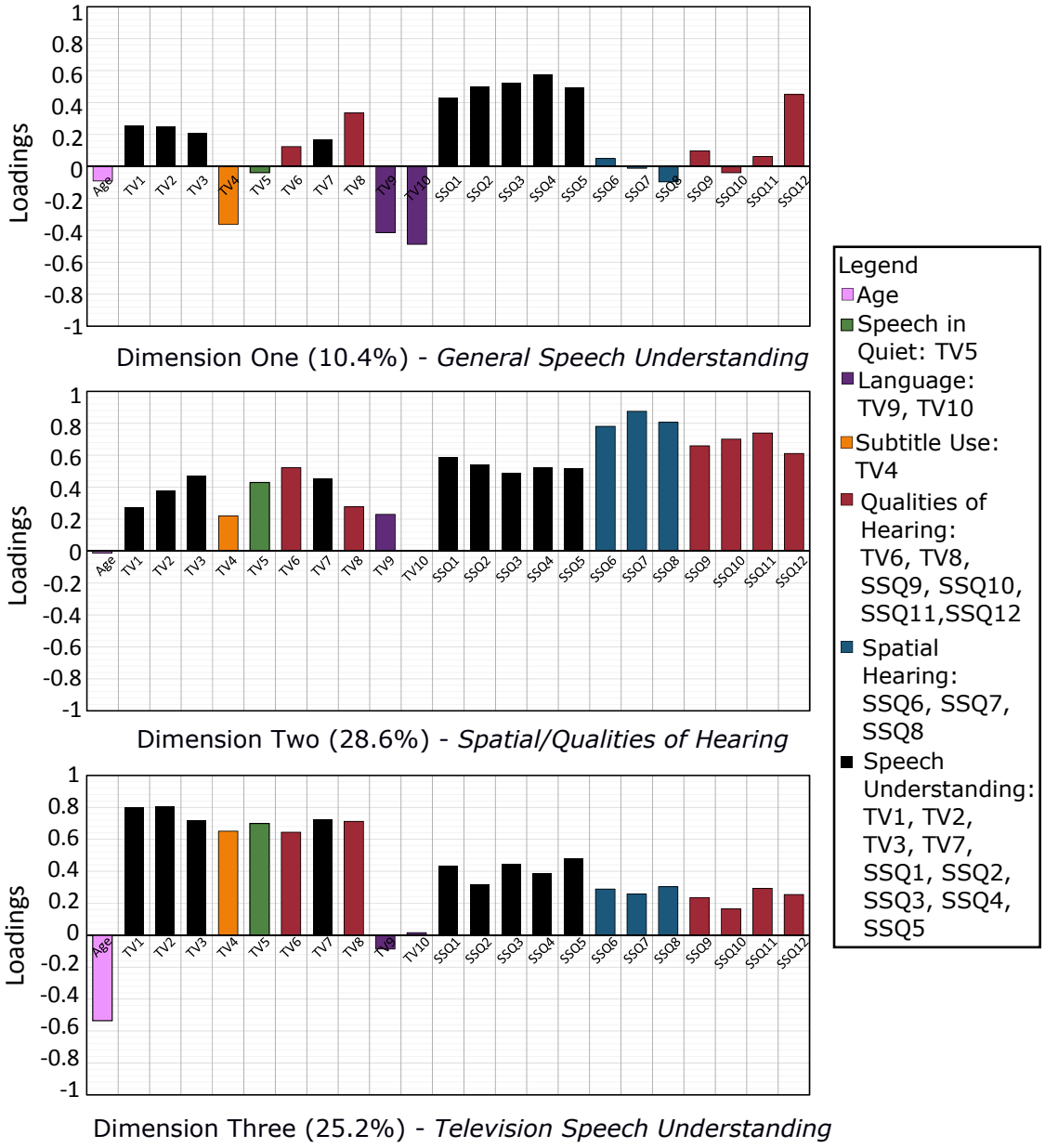


Figure 3.6 Loadings (correlation) between the first three principal components (PCs) after orthogonal rotation and the items from the reduced Speech, Spatial and Qualities of Hearing scale, the television experience questions from Table 3.2 and participants' age. It can be seen that Dimension One, named 'General Speech Understanding' is most strongly correlated with the speech understanding items. Dimension Two, named 'Spatial/Qualities of Hearing' is most strongly correlated with the spatial and qualities of hearing items items. Dimension Three, named 'Television Speech Understanding' is most strongly correlated with the television experience items.

The $\lambda > 1$ and inflexion point methods showed one significant dimension whilst parallel analysis indicated four. To further investigate which dimensions should be retained, rotation was performed on 2 to 4 PCs, the results of which can be seen in Table 3.8 (as well as the variance contained in the first un-rotated dimension). It can be seen that the addition of a fourth dimension does not add much explanatory power in the fourth dimension (5.9%). Three dimensions gives a good amount of explained variance in each dimension and offers greater overall explained variance than retaining only two dimensions. As such, three dimensions were selected.

The loadings of each of the items onto the first three dimensions can be seen in Figure 3.6. These are colour coded to aid interpretation; questions from the SSQ12 are colour coded based on the segment they come from in the original SSQ49 whilst the television version are categorised based on whether they refer to speech in noise, speech in quiet or qualities of hearing. Dimension One, which represents 10.4% of the variance, has a strong positive correlation with items about speech understanding and negatively correlated with subtitle use and language items (second language and British sign language viewing). Items from the *'Qualities of Hearing'* scale about effort and concentration are also weakly loaded onto this dimension. This dimension can be considered representative of *'General Speech Understanding'*.

The second dimension represents the largest amount of variance in the data: 28.6%. Most of the items have a positive correlation with this dimension, with the SSQ12 items loading more heavily onto this item than the television items. Of the SSQ12 items, this dimension is dominated by 'Spatial Hearing' and 'Qualities of Hearing'. As these factors were not as widely addressed in the television subscales as the other factors from the SSQ12, it is logical that the Spatial and Quality items together would represent a significant portion of the variance in the data. As such, this dimension is denoted as *'Spatial Hearing and Qualities of Hearing'*.

Dimension three represents 25.2% of the variance in the data. The third dimension is dominated by television specific questions and negatively correlated with age. The dimension is also weakly loaded by general speech understanding items from the SSQ12. 'Qualities of Hearing' items specific to television load heavily onto this dimension, whilst 'Qualities of Hearing' items from the SSQ12 load only very weakly. This supports the idea that this dimension is dominated by factors specific to television and is denoted as *'Television Speech Understanding'*.

3.5.3 Supplementary variables

Supplementary qualitative variables representative of demographic information and hearing ability were utilised to determine how different groups varied in the three new dimensions. For each group, the mean value and a 95% confidence interval around that mean are shown in addition to the projection of individual's scores onto the new dimension.

Device

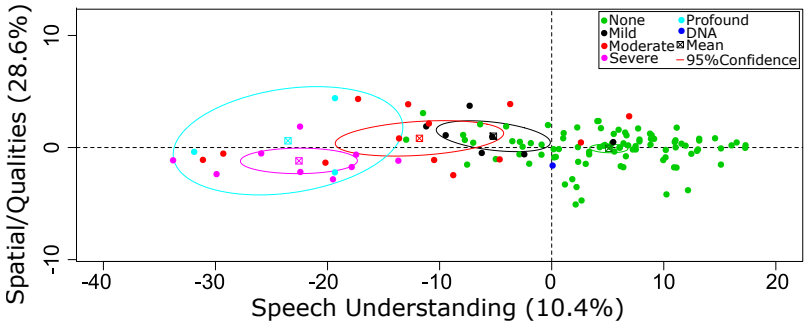
Respondents were asked to identify the device they were responding on, in order to ensure that some respondents did not give low responses on Likert scale due to presentation issues with the scale on their devices. Investigating group differences between respondents using PCs, phones and tablets it was seen that there was no significant difference between the responses for different device users, as such no data needed to be excluded on this basis.

Country of residence

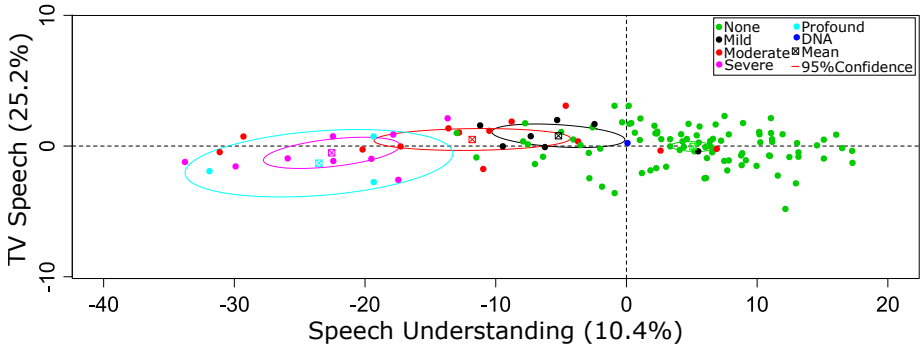
Country of residence was investigated and it was determined that there were no significant differences between individuals resident in countries other than the UK and those resident in the UK. Given this, and the similarity in identified problems with intelligibility and access in Germany and Japan outlined in Chapter 2, it was deemed that no data needed be excluded on this basis. The caveat on this result is that only 17.4% of the dataset identified countries of residence outside the UK, the majority of those being from Australia which has much in common with the UK in terms of demographics and television content. Should greater amounts of data be collected from residents of other countries, a significant difference between them may become apparent.

Severity of hearing impairment

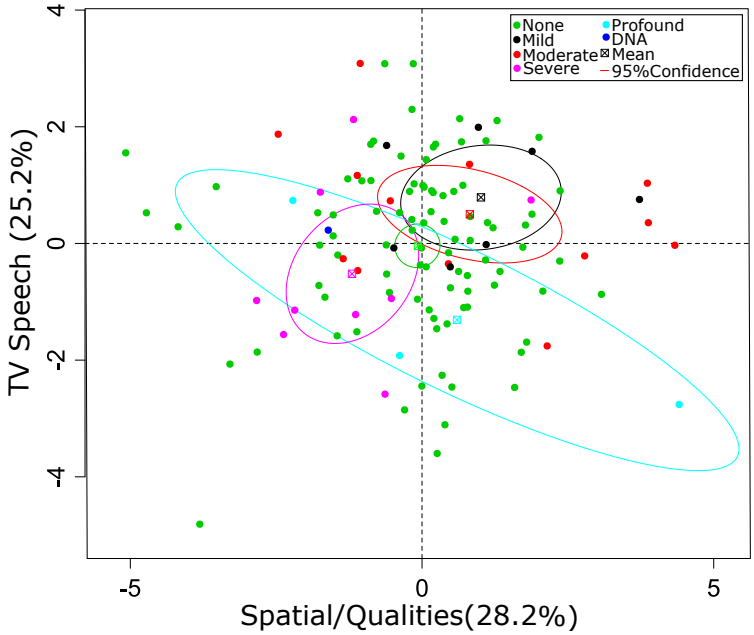
The results of grouping the individual projected scores by self-identified severity of hearing loss (or no loss) can be seen in Figure 3.7. There is a significant difference between the normal hearing group and all severities of hearing impairment along the dimension *'General Speech Understanding'*. Individuals score increasingly negatively on this dimension as their hearing loss becomes more severe. From Figure 3.7a, it can be seen that the mild and moderate loss groups score higher on the *'Spatial/Qualities of Hearing'* dimension, whilst severe loss score lower. Normal hearing respondents also score slightly lower on this dimension; for speech understanding, no trend from normal to increasingly severe hearing loss is evident. Mild and moderate hearing loss individuals score higher on *'Television Speech Understanding'* whilst



Individual participant scores projected onto the *General Speech Understanding* and *Spatial/Qualities of Hearing* dimensions.



Individual participant scores projected onto the *General Speech Understanding* and *Television Speech Understanding* dimensions.



(a) Individual participant scores projected for hearing loss severity onto the *Television Speech Understanding* and *Spatial/Qualities of Hearing.* dimensions

Figure 3.7 Mean and 95% confidence ellipses for the supplementary variable: Severity of Hearing Loss, as a function of all three dimensions. Those who did not answer this question are noted as 'DNA'.

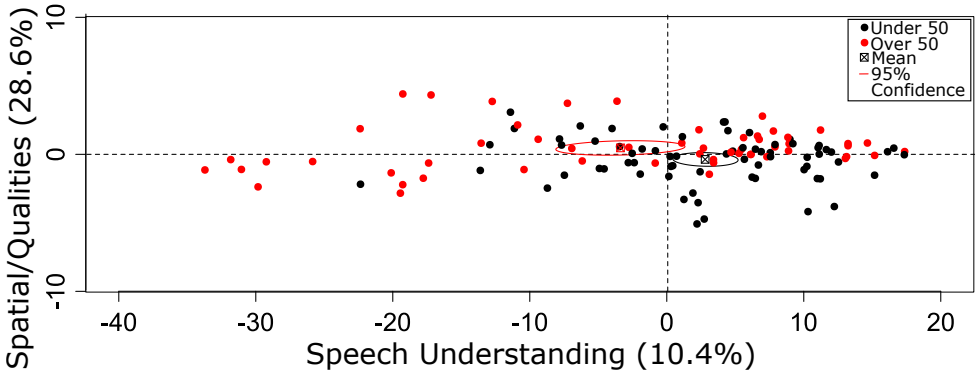
severe to profound losses score lower. This may be because for those with mild to moderate loss, TV speech provides a more controlled listening environment than day to day listening. However, we can seen that the mean for normal hearing listeners is located close to zero and spreads across the range of values of *'Television Speech Understanding'*. This highlights a key result which indicates that the variation shown in *'Television Speech Understanding'* cannot be explained by hearing loss alone. Other factors are at play in determining whether an individual has difficulty understanding television speech.

Age

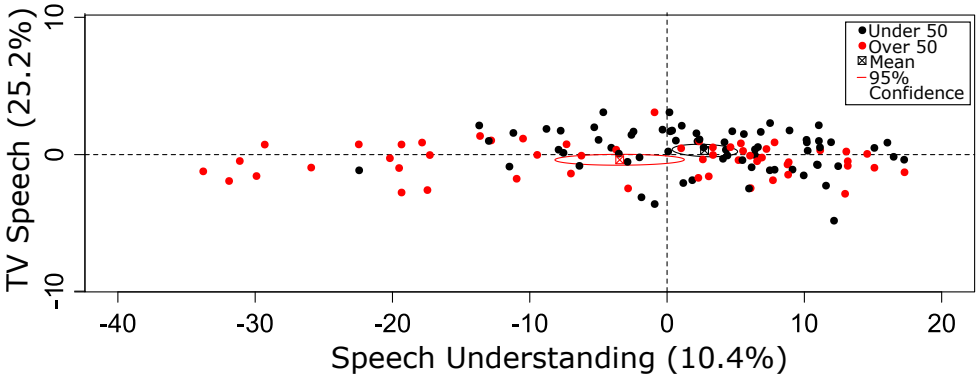
Respondents were segregated into older, defined here as over 50 years of age, and younger, between 18 and 49 years of age, to determine whether any differences were present due to age. It can be seen from Figure 3.8 that older respondents on average score lower on the *'General Speech Understanding'* dimension whilst younger listeners score higher. As the 95% confidence ellipses do not overlap, this difference can be considered significant. This mirrors the relationship seen for severity of hearing loss. This is to be expected as hearing loss increases with age and, additionally, speech in noise understanding often reduces with age independent of hearing loss [256]. Furthermore, on average younger listeners score higher onto the *'Television Speech Understanding'* dimension whilst older listeners score lower, indicating that regardless of hearing loss, older respondents have greater difficulty understanding speech on television.

Native language

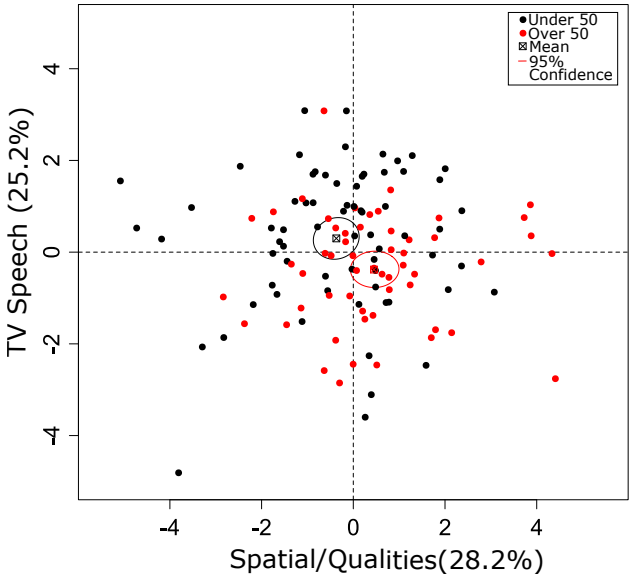
As the vast majority of respondents identified English as their native language, those who identified another native language were grouped together under 'other'. It can be seen in the Figures 3.9a and 3.9b that native English individuals score higher on the *'General Speech Understanding'* dimension whilst other language groups score lower. The 95% confidence intervals of these two groups overlap however, indicating that these differences are likely not significant. However given that the majority of respondents live in predominantly English speaking countries, it is to be expected that *'General Speech Understanding'* would be less for those who do not have English as their native tongue. There is no significant difference seen in Figure 3.9c, with the confidence ellipse of English speakers being completely encompassed by the confidence ellipse for other languages. This indicates whilst native language generally affects speech understanding, it doesn't impact on television speech understanding (potentially due to subtitle use and other strategies designed to aid second language viewing).



Individual participant scores projected onto the *General Speech Understanding* and *Spatial/Qualities of Hearing* dimensions.

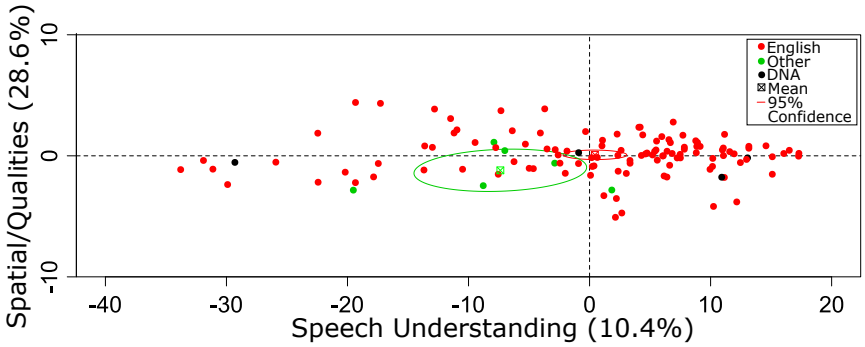


Individual participant scores projected onto the *General Speech Understanding* and *Television Speech Understanding* dimensions.

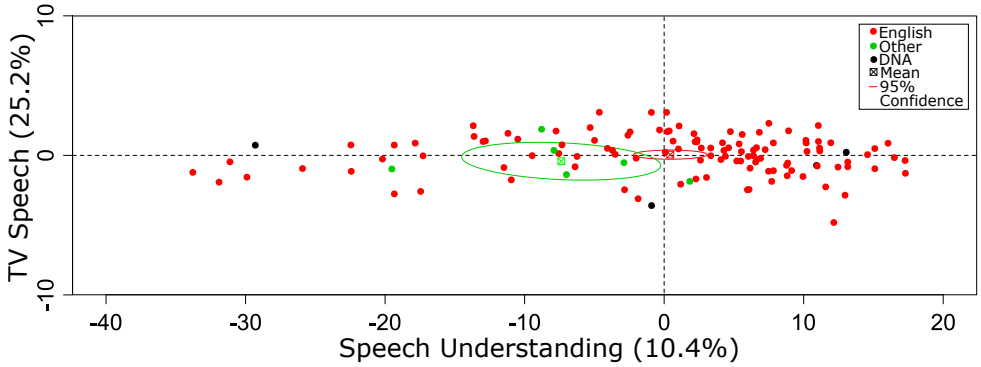


(a) Individual participant scores projected for hearing loss severity onto the *Television Speech Understanding* and *Spatial/Qualities of Hearing.* dimensions.

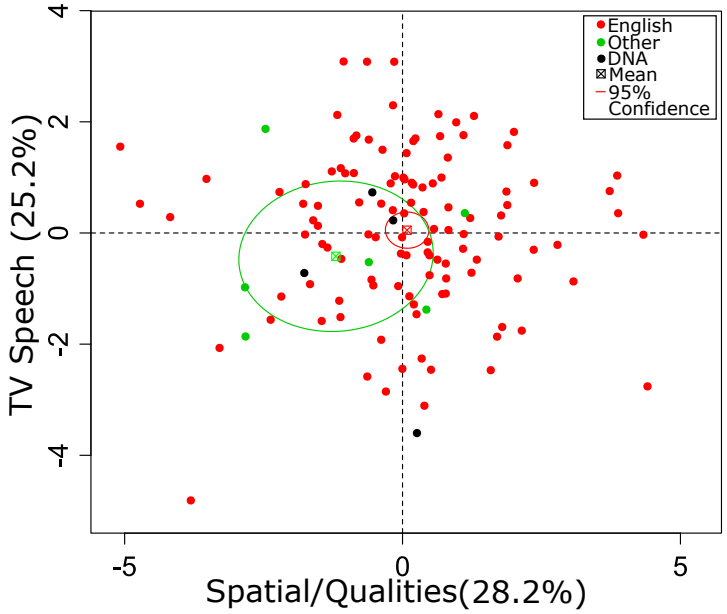
Figure 3.8 Mean and 95% confidence ellipses for the supplementary variable: Age, as a function of all three dimensions.



(a) Individual participant scores projected onto the *General Speech Understanding* and *Spatial/Qualities of Hearing* dimensions.



(b) Individual participant scores projected onto the *General Speech Understanding* and *Television Speech Understanding* dimensions.



(c) Individual participant scores projected for hearing loss severity onto the *Television Speech Understanding* and *Spatial/Qualities of Hearing* dimensions.

Figure 3.9 Mean and 95% confidence ellipses for the supplementary variable: Native Language, as a function of all three dimensions. Those who did not answer this question are noted as 'DNA'.

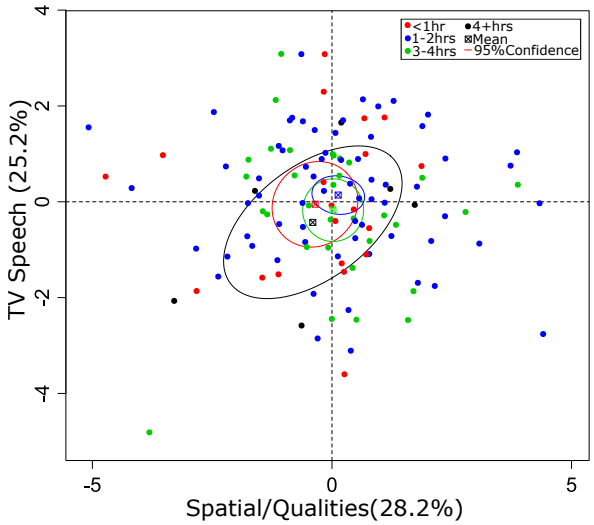


Figure 3.10 Individual participant scores projected for hearing loss severity onto the *Spatial/Qualities of Hearing* and *Television Speech Understanding* dimensions, with the mean and 95% confidence ellipses for the supplementary variable: Hours of Television Watched Daily.

Hours of television watched daily

Figure 3.10 shows individuals grouped by the amount of television they watch on average each day. It can be seen that there is no significant difference between the groups, even in the dimension *'Television Speech Understanding'*. This suggests that the amount of television viewed does not improve the understanding of speech on television. Alternatively this may indicate that poor understanding of speech on television does not alone deter individuals from television viewing.

Musicianship

All respondents indicated their level of musicianship. It can be seen from Figure 3.11 that amateur musicians score higher on the *'General Speech Understanding'* dimension whilst respondents who are not musicians score lower, to a similar degree as older listeners and those with mild hearing loss. Whilst there is still significant debate about the role of musicianship in speech perception [252, 253], these results suggest that musicianship does correlate with greater speech perception ability. There were only a small number of respondents who identified as professional musicians and so whether professional musicians score higher on speech understanding than amateur musicians cannot be determined from the available data. There was no significant difference in the *'Spatial/Qualities of Hearing'* or the *'Television Speech Understanding'* dimensions (and as such, these are not shown here).

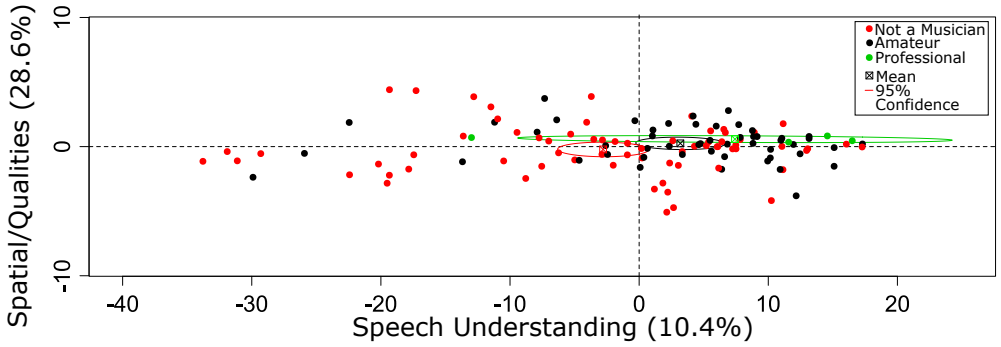


Figure 3.11 Individual participant scores projected for hearing loss severity onto the *General Speech Understanding* and *Spatial/Qualities of Hearing* dimensions, with the mean and 95% confidence ellipses for the supplementary variable: Musicianship.

3.5.4 Discussion

From the PCA, it can be seen that individuals can be characterised within a three dimensional space comprised the dimensions: *'General Speech Understanding'*, *'Spatial/Qualities of Hearing'* and *'Television Speech Understanding'*. The first two dimensions, which represent a combined 39.0% of the variance in the data and are interpreted here as the dimensions of *accessibility*, relate to an individual's sensory and cognitive media access needs. The third dimension, which represents 25.2% of the variance, is based primarily on the characteristics of television speech alone. Thus, it is termed here as the dimension of *audibility*. Furthermore, given the use of orthogonal rotation, these dimensions are uncorrelated indicating that elements of poor speech understanding owing to hearing loss are independent of speech understanding problems based on capture and delivery factors, such as those identified in Section 3.4.5. This is corroborated by Figure 3.7 which demonstrates that the difficulty (or ease) of speech understanding specifically for television cannot be explained by hearing (dis)ability alone.

The *accessibility* dimensions are primarily defined by hearing ability, however native language, age, and musicianship also effect these dimensions. It has been demonstrated that speech understanding in noise is easier in your native language than subsequent languages, which may explain the effect of native language on these dimensions [275]. However to conclude this would require further investigation with a larger population of non-native English speakers. It has been established that, even without audiometric hearing loss, age affects the ease of speech understanding in noise. Subsequently, it is logical that this affects scores on the *'General Speech Understanding'* dimension. Similarly, musical training is positively correlated with speech understanding in noise, as demonstrated by other studies

[252, 253]; combined, these factors likely form predictors for an individual end-user's accessibility needs.

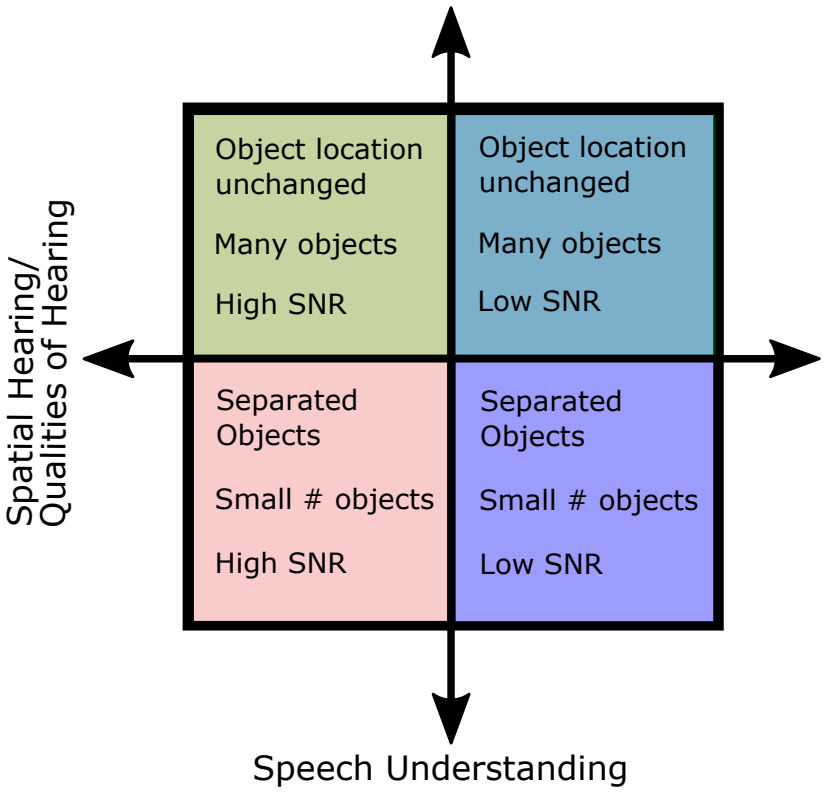
Age reliably predicts problems with speech understanding, both in television and in general terms. Given that speech understanding degrades with age, irrespective of hearing loss, this is to be expected. Particularly in the case of television speech, understanding may also be influenced by age-based differences in the type, genre, and origin of the programmes watched.

It is heartening to note that news and current affairs programmes remain the most intelligible, as their understanding is crucial to societal engagement. However, the high rate of problems with drama content is concerning, given its importance to communicating our shared culture as well as showcasing differing cultures and their stories. The problems caused by accent add to this and indicate that more needs to be done to ensure all viewers have access to cultural stories. This provides another example where, by being armed with knowledge of the problem, other measures can be deployed to compensate. Additional compensatory measures, for content which may be problematic, would include ensuring the quality and availability of subtitling for drama content or having the speaker's face in shot, to facilitate lip-reading.

Bringing together the problems identified in Section 3.4.5 and the new three dimensional space, we can delineate which problems relate to *audibility* of content and which relate to *accessibility*. Problems in programme capture, such as clarity of dialogue and choice of shots, likely form part of this independent *audibility* dimension. If speech is recorded poorly, this will affect all listeners. Balance between different objects and the level of added music can be considered to primarily exist in *accessibility* dimensions, as hearing loss significantly influences the required speech to background ratio required for understanding [170, 276]. The caveat is that, at extreme speech to background ratios, all listeners would be affected; this is also supported by the results in Figure 3.7a. The second component of the *accessibility* dimension includes spatial factors, though changes to spatial location of voices was not identified as a problem in Section 3.4.5. From patterns of subtitle use, it can be seen that usage is not exclusive to hard of hearing listeners and, as such, is likely influenced by all three dimensions. Similarly the quality of television speakers likely contributes to all three dimensions as poor reproduction will exacerbate problems with speech understanding, both if the listeners have some degree or hearing loss, or if the speech quality is initially poor.

Considering the identified areas of improvement within the new *accessibility* subspace, a corresponding subspace of solutions is presented in Figure 3.12. The *'General Speech Understanding'* dimension can be improved by greater speech levels as compared with other broadcast elements, or increased speech to background ratio. The *'Spatial/Qualities of*

Figure 3.12 Possible solutions to problems in the accessibility dimensions, projected on the *General Speech Understanding* and *Spatial/Qualities of Hearing* dimensions derived from the Principal Component Analysis in Section 3.5.



Hearing' dimension can be improved by having fewer auditory objects (to aid in auditory discrimination and cognitive load) and greater spatial separation between these objects.

3.6 Part I discussion: Scope of this doctoral work

This work aims to develop both novel understanding as well as new technology which translates academic knowledge into real, end-user benefit.

Utilising the literature reviewed in Chapter 2 and the characterisation of end-user needs completed in this Chapter 3, the focus of this doctoral work can now be defined. From Section 3.4.5 and Section 2.5 it evident that many individuals, particularly those with hearing impairment, struggle to understand speech in television. Furthermore, the level balance between different broadcast objects appears to be an area of significant concern, both shown in the end-user responses here in Section 3.4.5 and in previous studies summarised in Section 2.7.3. From Section 3.5.4, it is also apparent that ability to discriminate between objects (Qualities of Hearing) and their spatial location (Spatial Hearing) are significant defining factors of an individual's hearing abilities and should be considered in accessibility strategies. For these reasons, this work's broad focus will be on understanding how the balance between audio objects can be better adapted for different listener's needs.

From Section 3.4.3 and Figure 3.4, it can be seen that the most problematic genres are drama and film. These genres of content have the most complex soundscapes and it is likely for this reason that end-users find them the most challenging. These genres are also seen in Figure 3.1 to be the most watched genres of programming, therefore the area where the greatest positive impact can be made. Results from the survey indicate that foreground sounds, related to the narrative and action on screen, can be important to comprehension of the plot. This further supports the conclusions of Chapter 2 that there is significant potential to be explored in the personalisation of redundant non-speech sounds which is important to understanding. As such this work will specifically explore how the presence and balance of redundant non-speech audio objects affect television speech understanding, particularly for dramatic content.

Translating this new knowledge into end-user technology will be facilitated by advances in object-based audio technology. Given the variety of hearing impairments outlined in Section 2.2, it would not be likely that a single solution for all hard of hearing listeners could be found. Further to this, Section 3.4.5 indicates there is a desire from the end-user to have greater agency over how their content is reproduced and consumed in their own home. This combined need for adaptable solutions, and desire for agency, indicate that technology based on end-user personalisation is likely to be both effective and well received.

From the PCA performed in Section 3.5, it can be argued that the problems identified in Section 2.4 fall into two categories: *audibility* challenges and *accessibility* challenges. The majority of audibility challenges, such as poor delivery and capture, should be addressed from within the production and transmission chain. End-user technology and personalisation can only provide ad hoc compensation for it. The potential of object-based audio to deliver improvements in the accessibility dimensions is much more significant. Methods for end-user control of the balance between redundant non-speech audio objects and other objects for accessibility purposes will be the focus of the technology development portion of this doctoral work.

3.6.1 Research questions

The scope of this work is summarised in two research questions. They are:

- What is the relationship between redundant non-speech audio objects and broadcast speech intelligibility, for normal and hard of hearing listeners?
- Can a system be designed to allow end-users to control the balance between audio objects for dramatic content which is simple to use and preserves comprehension?

3.7 Chapter summary

This chapter has described the development, results, and analysis of a survey instrument designed to characterise end-user's accessibility needs. Development of the survey aimed to characterise respondents' hearing ability and subsequently investigate how this corresponded to their television viewing needs. It also investigated problems experienced by a variety of viewers. The results from implementing this survey were then analysed, utilising a mixed method approach.

The key conclusions from this chapter, which will shape the direction of this doctoral research, are:

- Drama and film are the most problematic genres, with news and current affairs programmes the most intelligible.
- Foreground sounds are considered the most useful non-speech audio object for understanding the plot of a drama.
- Three additions and one modification needed to be made to Armstrong's 'Audibility Problem Space'. The additions are: MULTIPLE SPEAKERS AT A TIME, BALANCE BETWEEN AUDIO ELEMENTS and USER CONTROL OF REPRODUCTION. QUALITY

AND AVAILABILITY OF CLOSED SUBTITLES needed to be modified to include the underlined text.

- Speech intelligibility problems in broadcast content stem from two separate and independent challenges: providing television speech which is sufficiently *audible* and providing *accessible* content.
- Given the variety of needs expressed by individuals, any approach to accessibility needs to be personalisable to be effective.

A solution space for hearing loss accessibility strategies has also been developed (seen in Figure 3.12), based on analysis of the survey data. From these results, the core research questions (Section 3.6.1) and the scope of this doctoral work (Section 3.6) were defined.

Part II

The Science

Chapter 4

Methodological approach for evaluating intelligibility: analysis and experimental design

4.1 Introduction to Part II

This part describes the design of the methodological approach for evaluating the first research question, outlined in Section 3.6.1:

What is the relationship between redundant non-speech audio objects and broadcast speech intelligibility, for normal and hard of hearing listeners?

This chapter first summarises the different methods for evaluating intelligibility used in other broadcast research. Methods are analysed with reference to three parameters: ecological validity, repeatability, and representativeness of the variable characteristics of hearing loss. From this analysis, the methodology for evaluating the first research question is defined. The developed method is then used to evaluate the relationship for a normal hearing cohort in Chapter 5. Initial investigation of the relationship for a hard of hearing cohort is then conducted in Chapter 6. Further development of the method is described in Chapter 7, followed by a final investigation with a hard of hearing cohort using the revised experimental method in Chapter 8.

4.2 Evaluating intelligibility

Investigating the first research question requires robust evaluation methods into the effect that non-speech sounds have on the intelligibility of broadcast speech. There is limited previous work specifically evaluating the relationship between non-speech sounds and speech intelligibility (see Section 2.2 for a review). For this reason, the following section conducts a more general review of intelligibility evaluation in broadcasting. This review is segmented into two parts: human measures and computational metrics.

These methods are then discussed with reference to three evaluation parameters: how representative they are of the varied characteristics of human hearing, their ecological validity, and their repeatability. The first of these evaluation parameters was selected because the characteristics of hearing loss are highly unique and vary greatly between individuals (as detailed in Section 2.2). The second of these parameters addresses the position put forth by Noble, that explorations of hearing loss should consider *auditory ecology*, not just the listener's hearing ability [277]. This is to say that evaluation of an individual's hearing performance should include not only the individual but also stimulus and an environment which is representative of the scenario under test. Finally, there is currently a widespread problem in the psychological sciences whereby some results are unable to be reproduced [278]. To ensure the rigour and reproducibility of this work, repeatability of measurements is selected as important evaluation parameter for any selected methodology.

4.3 Human intelligibility

Human intelligibility can be measured objectively utilising tests which determine how well an individual's speech perception performs under a particular set of conditions. It can also be measured subjectively by asking individuals to report their perceived level of intelligibility or speech understanding. This section summarises both objective and subjective approaches to human intelligibility evaluation.

4.3.1 Objective human measures

Human intelligibility is usually measured objectively by speech perception in noise tests. As outlined in Section 2.2, pure-tone audiometric thresholds do not accommodate the variability in individual's speech in noise performance [36, 256]. Speech perception in noise tests aim to capture this variability. There are a large variety of these tests including, but not limited to, the Hearing in Noise Test [279], Listening in Spatialised Noise-Sentences Test [280], Oldenburg Sentence Test [243], Bamford-Kowal-Bench Sentence Speech in Noise Test [281]

and the Revised Speech Perception in Noise Test (R-SPIN) [77, 82]. These tests take two main forms: everyday sentence tests using meaningful real-life sentences [82, 281] (e.g. the sentence *'The clown has a funny face'* [282]) and matrix sentence tests which follow a strict sentence form for each stimuli and require training [243] (e.g. *'Thomas wants nine cheap beds'* [283]). Regardless of type, these sentences are usually phonetically balanced across the stimuli and validated to ensure repeatable results. These tests often follow an adaptive paradigm which varies the speech to noise ratio until the 'Speech Reception Threshold', the point at which 50% of the speech is intelligible is determined [279, 284, 243]. An alternate approach is to utilise a static signal to noise ratio and quantify performance by the percentage of words correctly identified [285, 77]; the most commonly used maskers are multi-talker babble [77] or speech shaped noise [285].

Complementary intelligibility

Many speech perception in noise tests also facilitate the evaluation of complementary intelligibility factors as described in Section 2.3.2. The Listening in Spatialised Noise-Sentences Test accommodates for spatial release from masking [280] by using binaural reproduction of co-located and spatially separated speech and maskers to calculate the spatial advantage gained from separated sources. The effect of semantic context on intelligibility is quantified by the R-SPIN test [77, 82]. This is achieved through the use of high and low predictability everyday sentence stimuli where the speech preceding the keyword in these sentences either gives the listener clues to the keyword, e.g. *'Stir your coffee with a **spoon**'*, or no clues. e.g. *'Bob could have known about the **spoon**'* (where the keyword is noted in bold). The GRID corpus and it's extension, the Audio-Visual Lombard Speech Corpus, is an audio-visual corpus which has controlled audio-visual stimuli and matrix style sentences [240, 286]. The effect of static visual situational context, in the form of an illustration prior to the target sentence stimuli, is evaluated by the Illustrated Sentence Test [85, 287]. The stimuli following the illustration can be presented in an auditory-only modality to evaluate intelligibility or in a visual-only modality to evaluate lip-reading.

4.3.2 Subjective human measures

An alternate approach to objective word recognition scores is through self-reported ease of speech understanding. There are three main types of subjective measures which have been used in broadcast research: ratings, comparisons, and adjustment tests. The first is self-report ratings on intelligibility on a graduated or continuous scale. An example of an established and verified self-report measure of human speech understanding is the 'Speech,

Spatial and Qualities of Hearing Scale' (reviewed and used in Section 3.2.2) [249]. This asks participants to rate their hearing (dis)ability in different scenarios on an 11 point continuous scale. Other examples of subjective measures of intelligibility include comparisons where participants are asked to identify which of two or more stimuli they perceived as more intelligible [155, 123]. These may be blind (with a hidden reference) or the participant may be aware which signal is under test. These approaches may also be coupled with ratings of the level of improvement (or degradation) the preferred signal yields. The final type, adjustment tests, ask participants to vary certain parameters to a level which, for them, optimises the parameter under test [160, 174, 170]. This essentially requests participants to vary a system until it matches an external reference value or their own internal reference for a parameter, e.g. intelligibility [179]. These methods may be combined to form hybrid approaches, such as the 'Adjustment/Satisfaction Test' methodology developed by Fraunhofer IIS [180, 179] which provide flexibility and often facilitate the use of more ecologically valid stimuli. However, the generalisability of results from these studies is inherently limited by their subjectivity; without use of large, representative cohorts results only represent the average opinion of those involved in the experiment.

4.4 Computational intelligibility metrics

Computational metrics, often termed 'objective intelligibility metrics', are tools that produce proxy measures for human intelligibility judgements. These metrics are based on the effects that noise, signal processing, or a reproduction space may have on signal features. The way these metrics account for these effects fall into two main types [288]. The first type quantifies the *masked audibility* of speech in noise, utilising estimates of both the speech and noise to determine masking level. Examples of these include the Articulation Index [289], Speech Intelligibility Index (SII) [290] and the Glimpse Proportion [65]. In the case of the Glimpse Proportion, the masked audibility is evaluated by determining the number of masked spectro-temporal segments which are above a pre-set audibility criterion (often 3dB) [65]. The other type of metric quantifies the *distortion* caused by the masker by measuring the similarity between the clean and noisy speech [288]. This type of metric includes the Normalised-Covariance Measure [291], the Coherence Speech Intelligibility Index (an extension of the SII) [292, 293], the Modulation Spectrum Area [294] and the Short Term Objective Intelligibility metric (STOI) [295]. In the case of the Coherence Speech Intelligibility Index, the distortion is measured by determining the degree to which the noisy or modulated output of a system is linearly related to the clean input [292, 293]. The Speech Transmission Index (STI) is also a metric of the distortion type and calculates

the modulation transfer function of the transmission path [296]. It is most commonly used to evaluate the effect of room acoustics, rather than noise, on speech intelligibility. It has been suggested as suitable for speech in noise evaluations [75], though existing research contradicts it's suitability for this application [297]. The SII [290] and the STI [75] have been standardised. These metrics can either be intrusive, which require both the noisy speech and a clean reference signal [65, 290, 295, 291], or non-intrusive, which do not require a clean reference signal [298, 294].

Computational intelligibility metrics based on automatic speech recognition systems have seen renewed attention [299–302], as automatic speech recognition systems now match human recognition levels for some tasks [303]. One approach is to utilise the speech recognition algorithm to re-synthesise the clean reference signal required by intrusive metrics such as the STOI [301]; an alternate approach is to design speech recognition systems which use the same auditory models as humans [304]. A recent study has shown that the relevant features utilised by automatic speech recognition coincide with those used by humans, in particular the spectral peaks of the speech signal and dips in the noise masker [303]. Through shared auditory models, automatic speech recognition systems can be taught to make predictions of human intelligibility [304].

4.4.1 For hearing loss

Many intelligibility metrics have either been adapted for, or specifically developed to, model hearing loss. Humes demonstrated that the STI could predict hearing impaired performance by modelling elevated audiometric thresholds as internal noise [305]. The Modulation Spectrum Area metric, which developed from the STI's modulation transfer function, has been used directly as an intelligibility correlate for cochlear implant users in reverberant conditions [294]. The SII, though developed for normal hearing listeners, has also been used to predict speech reception thresholds in noise and different room acoustic conditions [306]. A number of adaptations to the SII have been explored to determine its efficacy for predicting hearing impaired listeners' speech reception thresholds in fluctuating noise. These incorporate hearing loss as an input parameter to the model, yielding small improvements to the original SII's predictions [307].

Metrics which have been specifically developed for use with hearing impaired listeners include the Hearing Aid Speech Quality and Perception Indicies (HASQI and HASPI respectively) [308, 309] and the Normalised Covariance Measure [291, 310]. Hearing loss is often modelled in objective metrics, through widening of filter bandwidths, to represent loss of frequency selectivity and attenuation of filter banks to represent loss of hearing sensitivity [311]. In the HASQI auditory model, which has been extended to form the basis

for predicting both intelligibility (HASPI) and quality (HASQI), loudness recruitment is also accounted for by a reduction in dynamic range compression [312]. The Normalised Covariance Measure is based upon a perceptual model which can accommodate different processing deficits including an individual's temporal forward masking ability [291].

Approaches based on speech recognition have also been designed to accommodate for hearing impairment [299, 300]. Fontan aims to develop a system which can predict intelligibility and comprehension test performance for hard of hearing listeners [300]. However their system has only been validated with normal hearing listeners and stimuli which has been altered to mimic varying degrees of presbycusis. The Framework for Auditory Discrimination Experiments (FADE), developed at the University of Oldenburg, has been shown to successfully predict the outcome of matrix sentence tests and psychoacoustical experiments [299]. The system utilises a front-end which models hearing impairment, utilising Plomp's Attenuation and Distortion model [38]. The *Attenuation* is modelled similarly to other metrics which reduce the level in each frequency band based on the audiogram [299]. The *Distortion*, or suprathreshold loss, is modelled as uncertainty by adding Gaussian white noise with a variable standard deviation. This variable value can be set based on a number of 'typical' audiogram shapes or can be determined with an independent data set.

4.4.2 For broadcast

Two objective intelligibility metrics have been developed for specific application to broadcast speech in recent years [174, 302]. The first was developed with the aim of providing a tool for broadcast production staff to evaluate the intelligibility of their mixes [174]. It is based on the Binaural Distortion Weighted Glimpse Proportion [313], a development from the original Glimpse Proportion [65]. This approach is double ended, requiring both the masked speech and original clean speech to evaluate the intelligibility. The second is a single ended approach which utilises an automatic speech recognition approach [302, 314]. It estimates the difficulty of perceiving the speech based on the distortion of the automatic speech recognition system's representations of each speech sound (posteriorgram). Both these approaches were developed and evaluated using a range of ecologically valid background noise including competing speech, music, and ambiences.

4.5 Intelligibility evaluation in broadcast research

This section gives an overview of the types of intelligibility evaluation methods which have been employed in previous broadcast research. It gives examples of each approach and is not exhaustive.

4.5.1 Computational metrics

A handful of researchers have made use of objective metrics in their studies. Müsch proposed that the SII could be used within a speech enhancement algorithm to evaluate intelligibility of 5.1 broadcast content but did not validate it [147]. Work by Mapp evaluated the effect which the home listening environment may have on the perceived intelligibility utilising the STI [104]. Whilst not a listener centred metric, Mapp leveraged the STI to show the majority (of tested) living rooms had good intelligibility and did not vary greatly. This suggested that the room effects are likely not a major cause of poor intelligibility of television and film content. The STOI metric has been investigated, among other performance and quality metrics, for detecting distortions incurred by dialogue enhancement systems [315].

In the ongoing development of the Binaural Distortion Weighted Glimpse Proportion [313] for measuring broadcast speech intelligibility, Tang et al. used it to evaluate the objective intelligibility of broadcast samples [174]. This evaluation was completed with young normal hearing, native British English speakers. It was shown that this measure predicts an optimal speech to background ratio which is on average 5.5dB lower than the preference level set by participants.

4.5.2 Human objective measures

The type of stimuli described in Section 4.3 has had some use in broadcast accessibility research. The 'DICTION Project' utilised the R-SPIN to evaluate their developed processor which removed background sounds in television for improved intelligibility [132]. They elicited both objective responses (target keywords) and subjective ratings of clarity and successfully evaluated the processor (though found it not to be effective for its stated aim). In the 'Clean Audio' project, the R-SPIN stimuli were adapted to evaluate whether the intelligibility of speech changed depending on whether it was presented via a central loudspeaker or a phantom centre [141].

More recent work has utilised the GRID corpus to evaluate the difference in the intelligibility with and without binaural externalisation of the speech [220]. As the intended application of this work is for binaural reproduction for listeners viewing on mobile devices,

GRID enabled the study to evaluate the effect of the presence of the visual of the speaker as well as improving the ecological validity of the study. Furthermore, the GRID corpus gives a variety of British accents which also makes the study more ecologically valid. Both speech shaped and speech modulated noise were used in this study.

Development studies to determine a relationship between the Binaural Distortion Weighted Glimpse Proportion and speech to background ratio by Tang et al. utilised the SCRIBE corpus [174]. This study used speech shaped noise, multi-talker babble, instrumental music, music with vocals, and competing speech; both studies utilised young, normal hearing participants.

Human objective measures have also been used in German and Japanese broadcasting research. In a follow-up experiment to earlier dialogue enhancement work by Fraunhofer ISS[162], Fuchs et al. utilised the Oldenburg Sentence Test to evaluate a dialogue enhancement strategy for hearing impaired listeners [203]. This implementation used both speech-shaped noise and 'applause' type noise and a cohort including both hard of hearing and normal hearing listeners. Some of the body of work from NHK has evaluated the intelligibility of their strategies. A very minimal early evaluation of their speech rate conversion system [134] evaluated by recognition of isolated Japanese syllables by three normal hearing females in their twenties [133]. Another study has used a speech in noise paradigm to evaluate optimal speech to background ratios [175]. This used mono Japanese syllables in pink noise. This was limited by the small cohort (9 elderly and 9 young listeners) and the fact that pink noise is not an ecologically valid background sound.

4.5.3 Human subjective measures

Ratings and comparisons

The earliest work in broadcast accessibility, conducted by Mathers and the BBC, utilised subjective ratings of quality for a variety of modified mixes [131]. Experiments within the 'Clean Audio Project' elicited subjective ratings of clarity of dialogue, sound quality, and enjoyment using blind A-B comparison [26]. Subjective ratings of quality have been used in many other works with varying degrees of success [131, 21, 148]. Subjective ratings of speech quality and general sound quality were also used by Fraunhofer in the 'Enhanced Digital Cinema Project' [148]. These studies all involved the target population, either hard of hearing or elderly, within their research process.

Work by Pearson used ratings of 'narrative clarity' mixed with a speech in noise paradigm to ensure listeners were presented with challenging but ecologically valid audio scenes [239]. In this work, participants listened to three different mixes of four audio-only scenes in multi-talker babble noise. For each scene, three mixes had been created and then presented

in noise: dialogue only, dialogue + key sound effects to the narrative (as determined by the producer), and the whole mix. Both naive and expert listeners were involved and all listeners were aged under 35. Results from this study were inconclusive.

Subjective ratings were also used in German language work by Hildebrandt where participants rated (using a questionnaire) their listening fatigue and the general pleasantness of TV excepts with different dialogue levels. [118]. This also utilised ecologically valid television content as well as reproduction through television speakers.

The use of subjective human measures by NHK in Japanese language work have varied in their rigour, cohort size, and representativeness. These evaluations range from subjective acceptability ratings of varying speech rates with representative (aged in 60s to 80s) cohorts of 356 people [136] through to evaluation with cohorts of six people or less [133, 139]. Expert listeners have been used, such as cohorts of active mix engineers, to determine the ranges of speech to background ratios acceptable in content production [175]. In contrast, a study evaluating acceptable loudness levels for elderly listeners have used normal hearing listeners ages 20-25 [139]. One particular study investigated the effectiveness of adapting programme level in spectral bands to the 'average' threshold losses in elderly listeners [137]. Whilst utilising representative cohorts, this study utilised the same subjects with which the 'average' had been created to evaluate the strategy [137]. This only proves the efficacy of the strategy for people for whose hearing characteristics the system was designed – the result is neither reproducible nor generalisable.

The first and second HHB4ALL studies utilised a comparison and rating approach where participants listened to both processed and unprocessed versions of the stimuli and were asked to identify how much better or worse they found the processed stimuli [123, 122, 155, 153, 156]. The study's design rightly recognises that hearing aid use will have an effect of the perception of any processing conducted on the broadcast signal and segregates participants into aided and unaided hard of hearing cohorts. However, the results were skewed by the use of a primarily severely hearing impaired cohort [123]. This undermines their claims that the developed algorithms were applicable for all hard of hearing listeners.

The environment of the studies was designed to have acoustics of 'living room quality' with the aim of ecological validity, as was the use of real audio-visual samples. It is not apparent whether the level of complementary cues (e.g. ability to lip-read) was controlled for in the stimuli. The studies were undertaken with up to six people participating concurrently. Whilst multi-person viewing presents an ecologically valid scenario for television viewing, listening position (such as whether the participant was in the front row, back row or had an off-axis listening position) was not controlled for or taken into account in the analysis. This was to the detriment of the repeatability and validity of these experiments.

The final part of the HHB4ALL clean audio work were four trial programmes. For each piece of programme content, participants noted on a timeline the sections of the content which they found unintelligible or problematic [156]. The original and processed versions were evaluated by different cohorts to ensure that effects of content learning were not introduced.

Adjustment

Object-based broadcasting's personalisation potential has been used to evaluate intelligibility and quality subjectively by allowing participants to adjust mixes to their preferred balance of audio objects [170, 180, 162, 163, 217, 160, 212, 24]. These adjustments are elicited by asking the participants to: optimise their understanding of the content [170], make the speech easy to follow [180] or set preferred balance with respect to the broadcast content [162, 163] or in light of background noise [217]. The studies in the Orpheus project utilised a different approach allowing participants to rate the perceived value of adjustment features before and after trialling them [24]. The majority of these studies used ecologically valid audio-visual material [170, 180, 212, 163, 24]. Many studies utilised self-selecting members of the general public [163, 212, 24]. Only one study recruited a hard of hearing cohort [170]. The common result from all these studies was significant interpersonal variation with normal hearing listeners often preferring higher levels of background noise to enhance the feeling of *'being there'* [163].

An adjustment based methodology for the evaluation of dialogue enhancement technology has been developed by Frauhofer IIC [180, 179]. This presents a similar first step to other adjustment methods where the dialogue is varied to a level relative to the background which optimises intelligibility and quality. The additional step that this approach introduces is an evaluation of the satisfaction after the level has been set which gives participants the opportunity to evaluate their selection with reference to the original and rate how much better (or worse) they perceive their setting to be. This allows evaluation of not only how personalisation features would be used (adjustment step) but provide quantitative evaluation of how those features affect quality of experience. Furthermore, if the original stimuli were selected as better in the satisfaction stage, this provided a post-test criterion to determine whether the task was understood and completed correctly. This improves the rigour and reliability of the results.

Other subjective approaches

Other subjective approaches have been used which do not directly assess intelligibility. These include use of focus groups to discuss desired features [24, 238] and production trials

[21, 24, 212]. Textual analysis has also been used to identify sources of intelligibility issues, such as in the analysis of complaints to German broadcaster ARD by Hildebrandt [118]. These approaches complement objective and subjective evaluations of intelligibility but are not sufficient approaches on their own.

4.6 Comparison of methods

The above methodologies are compared here with reference to the three evaluation parameters: representativeness of hearing loss, repeatability, and ecological validity. Based on this, selection of a methodology has been made for investigating the first research question.

4.6.1 Representative of hearing loss characteristics

Objective human measures such as speech in noise tests inherently account for the effect of suprathreshold hearing factors and tests of this type have been used to good effect in broadcast research [132, 141, 203, 174, 220]. Whilst some research has used a single speech to background ratio for all listeners for each stimuli type [203], the approaches used in the 'DICTION Project' [132] adjusted the signal to background ratio for each hard of hearing participant. This approach limits the generalisability of the results. Speech in noise tests designed to yield a speech reception threshold, like the Hearing in Noise Test [279], or modified versions of single speech to noise ratio tests, such as the multiple signal to noise ratio version of the R-SPIN, would likely accommodate for various hearing abilities more effectively [190].

Subjective human measures also inherently account for the various characteristics of hearing loss through capturing the lived experience of the cohorts included in the studies. This relies though on the cohorts used being representative. Approaches like Fontan's, where validation is undertaken with normal hearing listeners and stimuli modified to mimic hearing loss [300] will not be fully representative of hard of hearing listener performance. They are limited in two ways; the modelling of hearing loss used to adjust the stimuli is unlikely to model all suprathreshold factors and, more importantly, any mechanisms listeners have developed to compensate for their hearing loss, such as increased used of complementary intelligibility cues, will not be present in normal hearing populations.

Many studies described here have been undertaken with representative cohorts [25, 132, 170, 203, 118, 123, 155, 136, 139]. Some cohorts are very limited however with six or fewer participants, which prevents these studies yielding any generalisable results [139]. However, by utilising an experimental design which characterises participant variations (e.g. utilising

measures of hearing loss [170]) and only evaluating a single effect, small cohorts can be leveraged to give useful and significant data.

Whilst objective intelligibility metrics can be quite easily modified to accommodate attentuation characteristics of hearing loss (reductions in audiometric thresholds), the degree to which they can model suprathreshold loss varies [311, 308, 309, 291, 310]. The use of widened filters to represent loss of frequency selectivity begins to accommodate for basic suprathreshold factors and has the advantage of simplicity. However the greater the number of internal parameters the models contain, the greater the difficulty in calibrating the model and the greater the potential for the inaccuracies in the model to be larger than the difference between approaches under evaluation [37].

4.6.2 Ecologically valid

Many adjustment based methodologies [170, 162, 217, 180, 163] are highly ecologically valid as they present personalisation methods (such that the user may have available in an object-based broadcast system). Furthermore, through utilising real audio-visual media, these tests best accommodate for the inherent complementary intelligibility factors within television content. Similarly the majority of tests using subjective human measures have high levels of ecological validity through their content and are often conducted in ecologically valid environments [170, 155]. Use of commercial broadcast equipment, such as in built television speakers for reproduction of the audio [118] also increases the ecologically validity of the results.

The ecological validity of objective human measures vary depending on the type of sentences and noise used. Everyday sentence tests like the R-SPIN [82, 77], as compared with matrix tests, provide sentence stimuli which have greater similarity to television dialogue. Multi-talker babble noise also adequately represents many common problematic background sounds within television content, such as the ambience in crowded places like restaurants or shows recorded in front of live audiences. This use of everyday sentences and representative noise likely contributes to R-SPIN's use in television research [132, 141, 203]. Fuchs' use of the Oldenburg Sentence Test [243] in applause type noise [203] shows another adaptation which balances ecological validity with controlled stimuli. Usage of noises potentially encountered in the listening environment like the instrumental music, music with vocals, and competing speech used in Tang et al.'s evaluation of his broadcast intelligibility metric [174] may also improve the ecological validity of speech in noise tests. However, even with relevant noises, such tests remain limited by their audio-only modality. The Illustrated Sentence Test, with static visual context, goes some way to amending this [287]. The GRID corpus [240] improves on this further with matched vision of the speaker allowing for the

effect of visual complementary intelligibility cues, such as lip-reading, to be accounted for. However, as the speaker is face direct on to the camera throughout, it would only generalise well to content like news broadcasts and not to most dramatic content.

Objective metrics cannot account for complementary intelligibility cues within television content. hHwever they can be effective for evaluating the effect of background sounds on signal-dependent intelligibility [174]. Furthermore their ecological validity is heavily dependent on their efficacy across different broadcast noise types. Studies have shown that whilst state-of-the-art metrics make good estimates of intelligibility in single masker types, across-noise predictions are generally poor [288]. As such, without evaluation in different broadcast noise types, the ecological validity of most objective measures remains unknown and would require systematic reviews of existing methods in broadcast type noise.

4.6.3 Repeatable

Objective intelligibility metrics are, by their nature, repeatable which is one of their distinct advantages as they utilise static models of intelligibility. Most speech in noise tests have been validated with the target population also giving them good repeatability. However, this repeatability may be limited, if significant alterations are made to the stimuli or test methodology or if the stimuli are outdated (either in recording quality or sentence contents).

Subjective measures have some repeatability limitations due to the acquisition of personal preferences and self-reported data. Rigorous characterisation of personal and external factors which may influence or interact with these subjective judgements are necessary to ensure the validity of this type of methodology; the majority of described studies have not collected this additional data. The repeatibility and validity of these types of measures are improved by hybrid approaches combining adjustments and ratings of improvement (as in the Adjustment/Satisfaction test [180, 179]).

An aside on preference

In addition to the evaluation parameters considered here, preference must be considered in accessibility research; what is objectively better may not always be what is preferential to the target users. Taking an example from the established field of subtitle research, it has been recommended that subtitles should not exceed objective measured maximum reading speeds however target users prefer that subtitles are time-aligned, even if they exceed this speed [88, 316]. Identification of these preferences is only possible through methods which are both guided by the target population as well as involve them as research participants. So whilst objective measures and computational metrics provide useful, repeatable and often

representative evaluation methods, only working directly in consultation with end-users can we ensure ecological validity. This approach is taken in Part III where the second research question, which aims to develop end-user accessibility tools, is addressed.

4.7 Experimental design

It is evident from the above comparison that no singular methodology meets all criteria. Subsequently, mixed methodological approaches, combining the use of one or more methods, are likely to give the greatest reliability.

As the first research question Section 3.6.1 addresses an area of research with limited previous exploration, the results of this initial study should aim to prioritise repeatability and generalisability over ecological validity to specific aspects of broadcast viewing. To ensure that a fundamental understanding of the psychoacoustic phenomenon is gained, the experiment should be undertaken with cohorts of both normal and hard of hearing individuals. To yield significant and repeatable results even with small cohorts, an experimental design with limited factors under test should be used.

As the basis for the design of this experiment, the R-SPIN test was selected. This ensured that the experiment captured the effects of complementary intelligibility cues and the characteristics of hearing loss. Furthermore through the R-SPIN's use of multi-talker babble noise and everyday sentences, the test balances ecological validity of the stimuli with repeatability.

The R-SPIN stimuli consist of short, phonetically balanced sentences spoken by a male speaker of American English. All sentences end with a monosyllabic noun, the keyword, which participants are scored on their ability to correctly identify. The R-SPIN provides an objective measure evaluating both top-down and bottom-up processes involved in understanding speech in noise [77]. The original authors achieved this by controlling the predictability of the sentences, either giving the listeners clues to the keyword (e.g. *'Stir your coffee with a **spoon**'* and termed *high predictability* - **HP**) or no clues (e.g. *'Bob could have known about the **spoon**'* and termed *low predictability* - **LP**). These are conditions LP and HP in Table 4.1. Recognition of the keyword in these LP sentences relies entirely on receiving the acoustic signal of the keyword correctly. The HP stimuli differ in that the surrounding sentences allow for the use of top-down processing; any ambiguity in the keyword's acoustic signal can be resolved using knowledge of the English language and the contextual information provided by the sentence.

The original R-SPIN is a single factor experiment evaluating the effect of predictability on speech understanding. To address the first research question of this work (Section 3.6.1),

a second factor was added – redundant non-speech sounds conveyed by sound effects (SFX). These were added to both low and high predictability sentences and are labelled as conditions LP+SFX and HP+SFX in Table 4.1. The SFX selected were taken from broadcast quality SFX libraries (BBC Sound Effects Library [317] and Soundsnap [318]) for ecological validity. They were selected to give equivalent priming effect as the semantic cues in the HP sentences; for example, the HP sentence for the keyword *pet* is *'My son has a dog for a pet'*, which utilises the assumed knowledge that children often keep pet dogs. As such, the SFX selected for this keyword was a dog's bark. The SFX selected was always based on the HP version of the sentence and its usage of the keyword. The same SFX was used for both HP and LP versions of the sentence for comparability. The SFX used were not limited to recordings of the keyword itself but also used combinations of sounds, e.g. the SFX for the keyword *pond* consisted of both the sounds of water splashing and ducks quacking.

Table 4.1 Conditions for experimental methodology using modified Revised Speech Perception in Noise Test (R-SPIN [319], noting the number of stimuli and presence of redundant non-speech sounds for each condition.

Condition	Factor 1	Factor 2	# Stimuli
LP	Low Predictability	No SFX	50
HP	High Predictability	No SFX	50
LP+SFX	Low Predictability	SFX	50
HP+SFX	High Predictability	SFX	50

The 'Revised' aspect of R-SPIN, refers to work done by Bilger in 1984 to ensure that the test gave balanced performance for hard of hearing listeners. In doing so, he removed two lists and redistributed the remaining sentences using the psychometric data from 128 elderly listeners with sensorineural hearing loss. Validation of the new lists was performed with 32 of the original listeners who had a mean performance of 76% and 37% for the HP and LP sentences, respectively, at 8 dB speech to noise ratio. Utilising a modified version of the R-SPIN test, that also retains the original condition, allows for the results to be compared to previous implementations. This gives an indication whether the validation performed for the original experiment will likely hold.

A computational metric was also selected to complement the use of an objective human measure. The main role for this metric was to allow comparisons of the stimuli between different conditions in the experiment as well as between different studies in this work. Computational metrics are only proxy measures for the many and complex processes which occur in human speech recognition. This means any selected metric will have some limitations. To minimise the effect of these limitations, numerous works have conducted comparisons of common metrics' suitability for different types of masker and target signals [174, 288, 320, 321].

Guided by these works, the computational metric used in this work was selected based specifically on it's suitability for the R-SPIN masker and target signals and the experimental cohorts. The planned experimental work includes cohorts of both normal and hard of hearing listeners. Whilst there are numerous intelligibility metrics adapted to include hearing loss characteristics, these metrics run the risk of the models having inaccuracies which have a greater impact than the effect under test [37]. With this in mind, and considering that the stimuli would be consistent across the cohorts, a measure designed for normal hearing listeners was selected.

The R-SPIN has clean speech masked with multi-talker babble and the speech consists of short sentences and keywords. Given the use of clean, not distorted, speech a metric based on masked audibility rather than distortion would be more suitable. The R-SPIN's babble masker means that the selected metric must perform well in fluctuating noise and consider the temporal effects of the masker.

The most suitable masked audibility metrics are the Speech Intelligibility Index (SII) and the Glimpse Proportion (GP). The SII has an the advantage over the GP in that it is a standardised measure [290]. However, as it is a power-spectrum only model, it is limited in its usefulness for fluctuating maskers [288]. Debate surrounds the degree to which the GP accurately reflects the human processes involved in speech in noise perception, and whether the phenomenon of 'dip-listening' is valid [322]. Despite this, it demonstrates good correlation with human speech in noise performance [174] and outperforms the SII for fluctuating masker types [288]. Furthermore, the GP's quantification of short-term speech to background ratio is more suitable, given the short duration of the keywords in the R-SPIN. For these reasons the GP was selected to quantify and compare the signal-level intelligibility of the stimuli[1].

4.8 Chapter summary

This chapter has outlined the three main approaches to evaluating intelligibility which have been utilised in broadcast research: subjective human measures, objective human measures, and computational metrics. Their use in broadcast research has been outlined and then evaluated with reference to three parameters: representativeness of hearing loss, repeatability, and ecological validity. A methodology for addressing the first research question has been designed.

From this chapter we have concluded that:

[1]The adaption of the Glimpse Proportion for broadcast was not completed when the work in Chapter 5 was undertaken and, as such, that version was not used.

- Objective human measures best account for the idiosyncrasies of hearing loss and can be ecologically valid depending on design.
- Computational measures provide repeatibility but at the cost of simplifying the complexity of human hearing.
- Achieving ecological validity with subjective human measures is easy, however it requires representative and often large cohorts to ensure significant and meaningful results.

For these reasons, an objective human measure based on the Revised Speech Perception in Noise test was selected for investigating the first research question. However, it is evident from these conclusions that no single approach to evaluating intelligibility meets all the evaluation requirements and that the most reliable experimental design will likely be a hybrid approach. The computational metric, the Glimpse Proportion [65], was therefore selected to complement the use of the R-SPIN in the following chapters.

Chapter 5

Evaluating the relationship between redundant non-speech audio objects and broadcast speech intelligibility for a normal hearing cohort

5.1 Introduction

Utilising the methodology selected in Chapter 4, this chapter begins experimental work addressing the first research question outlined in Section 3.6.1:

What is the relationship between redundant non-speech audio objects and broadcast speech intelligibility, for normal and hard of hearing listeners?

A mixed methodological approach using a modified version of the Revised Speech Perception in Noise test (R-SPIN) [82, 77] and the computational metric the Glimpse Proportion (GP) [65] was selected. This chapter describes the methods' implementation with a normal hearing cohort under two conditions: where the non-speech sound and the preceding speech are overlapped in the time domain (Section 5.2) and when they are separated in the time domain (Section 5.3). Results from both a perceptual intelligibility experiment and an objective intelligibility analysis of the stimuli are reported for both conditions. Finally an empirical model was developed to describe the effect which the positioning of the non-speech object, relative to the target speech, had on the likelihood of correctly identifying a word in noise.

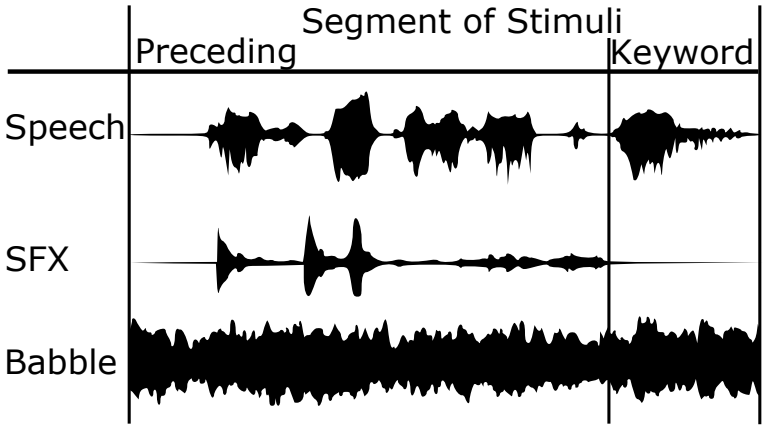


Figure 5.1 Example stimuli for Study One, showing the alignment of the redundant, non-speech audio object (SFX) with references to the preceding speech and keyword. Note that the SFX overlaps the speech preceding the keyword.

5.2 Study One: Effect of overlapping, preceding SFX on keyword intelligibility

The aim of this experiment was to evaluate the effect of overlapping, preceding redundant non-speech audio objects (referred to as sound effects, or SFX in this chapter) on keyword recognition. To achieve this, a modified version of the Revised Speech Perception in Noise (R-SPIN) test was employed [82, 77] with a second factor – non-speech cues conveyed by SFX. These SFX were introduced into half of the presented sentences. All SFX ended prior to the keyword being spoken (as seen in Figure 5.1) to ensure both types of cues had equal opportunity to prime the listener.

The conditions for the experiment can be seen in Figure 4.1, in Chapter 4. The original test segregated the 400 stimuli into 8 lists of 50. Four of these were selected; lists 1, 2, 5, and 6 constituting 200 sentences. In these 200 sentences, each keyword was presented twice, once with semantic cues and once without, giving 100 unique keywords. Of these keywords, 50 are shared between conditions LP and HP and 50 are shared between conditions LP+SFX and HP+SFX. Usage of different keywords in each condition pair was designed to reduce learning effects. Some sentences were swapped with those from other lists to ensure half the keywords could be paired with suitable SFX, however as much as possible of the original list integrity was maintained. The modified R-SPIN lists can be found in Appendix C.

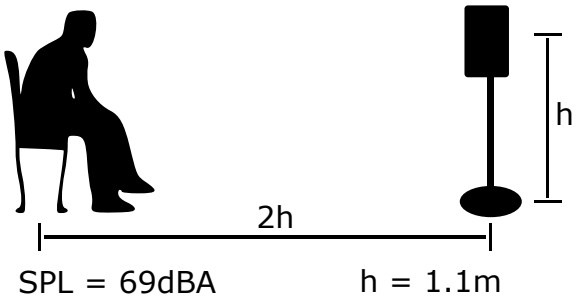


Figure 5.2 Experimental set up for Study One including participant location with reference to the loudspeaker, and height of the loudspeaker noted. Reproduction level set at the listening position is also given.

5.2.1 Implementation

The original test's single speech to background ratio (SBR) paradigm was maintained in this experiment [82]. It is not possible for all the SFX to have equal loudness due to their differing spectro-temporal qualities. In order to minimise the effects of this, all SFX were normalised to -23dB^{LKFS} over their duration before being mixed into the stimuli. This used as using a loudness meter meeting ITU-R BS.1770-2 specifications [323] and made the perceptual loudness of the SFX roughly equal.

Similarly, the speech, babble, and babble+SFX clips were all normalised to -23dB^{LKFS} over their duration. -23dB^{LKFS} was selected as the target level as this is the standard level for broadcast audio in the U.K. [95].

To set an appropriate SBR, which would yield a speech reception threshold of approximately 50% for conditions LP and HP from the original test, a pilot study was undertaken with experienced listeners ($n = 4$). Experienced listeners were defined as individuals who had substantial experience participating in listening tests From this an SBR of -2dB was selected, providing an average word recognition rate (WRR) of 53.5% across conditions LP and HP. The pilot study was also used to verify that any effects on WRR caused by the inclusion of the redundant non-speech audio objects in conditions LP+SFX and HP+SFX did not result in floor or ceiling effects [77]. Results showed a WRR of 72.5% and 80.0% for condition LP+SFX and HP+SFX respectively, avoiding ceiling effects.

The experimental set up can be seen in Figure 5.2. The sentences and babble+SFX mixture were co-located and presented from a Genelec 8030A Studio Monitor. The study was undertaken in a listening room meeting the ITU-R BS.1116-1 standard for listening tests [324]. This ensured that the environment was suitably representative of a home living room. The stimuli were presented at an SPL of 69dB(A) calibrated using pink noise, measured at the listening position. Participants were presented with 15 practice stimuli to allow them to

develop familiarity with the task and the speaker's voice. All participants were presented with the stimuli in pseudo-randomised order and the presentation order was counterbalanced across the participant pool to avoid learning effects.

Cohort

The cohort was made up of 24 native English speakers (7 females and 17 males), not inclusive of pilot participants, who had self-reported normal hearing. 13 participants were aged $18-29$ yrs, 7 were aged $30-39$ yrs and 4 were aged 40 or older.

Participants were asked how many previous research listening studies they had been a part of, to determine an approximately how much experience they had with this type of study. This was done as familiarity with this study of study may result in better performance, through understanding of the study's possible structure and aims. Participants were grouped into three levels of experience:

- Naive listeners: participated in < 5 studies
- Moderate experience: participated in > 5 and < 15 studies
- Experienced: participated in > 15 studies

The cohort had a variety of experience level, with half being naive listeners, 5 with moderate experience and 7 being experienced listeners.

5.2.2 Subjective results

Figure 5.3 shows the mean WRR with standard error bars. Condition LP has the lowest mean WRR of 35.8%. Condition HP gives a 73.5% improvement in recognition relative to condition LP, increasing the WRR to 62.1%. This result is consistent with other implementations of the R-SPIN test [190, 85]. For condition LP+SFX the WRR increased to 60.7%. This improvement of 69.5%, given it's similarity in magnitude to the improvement in condition HP, suggests that the complementary intelligibility cues offered by SFX yield a similar level of benefit for intelligibility as semantic cues. Condition HP+SFX shows a WRR of 73.7%, an improvement of 106.0% from the control condition and a 21.5% and 18.7% improvement from the conditions with only SFX cues and semantic cues respectively. This indicates that the combined effect of the cues also yields a modest improvement in intelligibility compared with either cue in isolation.

These results are lower than observed in the pilot study in Section 5.2.1. It is likely that this difference is due all those participating in the pilot having extensive listening study experience, whilst the experiences of those participating in the main study were varied.

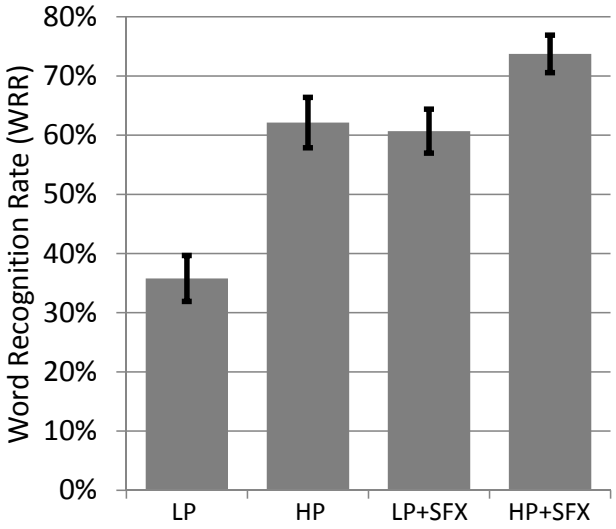


Figure 5.3 Mean word recognition rate ($n = 24$) for each experimental condition in Study One shown with standard error bars and where LP indicates low predictability sentences, HP indicates high sentences and +SFX indicates the inclusion of relevant, non-speech sounds.

Statistical analysis of subjective results

The subjective results are reported in Section 5.2.2 as aggregated WRR rates for ease of interpretation. However, to ensure that the greatest degree of accuracy is maintained in the statistical analysis, the individual dichotomous outputs for each participant's individual trials are analysed. This approach means that use of traditional statistical tools like the ANOVA were not possible. Standard methods for dichotomous outcome variables and predictors, such as logistic regression, also could not be used as the assumption of the independence of errors is violated by repeated measures design.

Instead the perceptual results were modelled using Generalised Estimating Equations (GEE) to determine the significant effects of the experimental factors and determine the size of these effects. GEE, seen in Eqn. 5.1, follows a similar form to logistic regression but additionally utilises robust standard error estimates to account for random or repeated factors. The GEE models were developed in R using the package geepack [325]. The data was tested for multicollinearity and complete separation; it was found not to violate these assumptions.

$$\text{logit}(\mu) = x_n^T \beta \tag{5.1}$$

where n is the number of factors

In addition to the two design factors used in Study One (SFX and PREDICTABILITY), four further categorical predictors were investigated (PARTICIPANT, EXPERIENCE, LIST and ORDER), which can be seen in Table 5.1. A saturated model was developed and tested using

Table 5.1 Design and additional factors investigated using Generalised Estimating Equation, which study they were used in and whether the factors were considered random tabled.

ID	Factors	Study: One	Two	Overall
Design Factors				
SFX	Inclusion of SFX	✓	✓	
PREDICTABILITY	High and Low Predictability	✓	✓	✓
OVERLAP	Whether the SFX overlaps or is separate from the speech (two levels)			✓
Additional Factors				
PARTICIPANT	Participant ID number	✓*	✓*	✓*
EXPERIENCE	How many previous listening tests they have participated in (3 levels: 1 - <5, 2 = 5-15, 3 - 15+)	✓	✓	✓*
LIST	Groupings of sentences (four levels, as in Appendix C)	✓	✓	✓*
ORDER	Presentation Order of the Lists (four levels)	✓	✓	✓
BOTH	Participation in both parts (two levels)		✓	✓*

$*$ denotes factors which were included into the model as random

Wald's test to determine which predictors offered significant improvements to the power of the model. The estimated β for the significant factors and their robust standard error, as well as the odds ratios, are shown in Table 5.2.

It can be seen from Table 5.2 that the design factors both have a highly significant effect on the model, with the effect of predictability being more than double that of SFX. ORDER was also found to be significant indicating a learning effect existed; however, given the odds ratio of 1.11, this effect is only small. Notably the interaction between SFX and PREDICTABILITY is significant indicating that the effects of semantic and non-speech cues were not strictly additive. One of the factors with the largest effect size was the interaction between PREDICTABILITY and EXPERIENCE. This suggests that their familiarity with general listening test structure may have allowed these listeners to better to make better use of the high predictability cues. However, whilst the predictor has a large odds ratio, it should be noted that this interaction is only weakly significant. There are also a number of weakly and moderately significant interaction effects, predominantly with odds ratios close to one, indicating their effect on the final result is minimal. The only highly significant interaction is between SFX, PREDICTABILITY and LIST. This indicates that performance across the different lists were not equivalent, suggesting that the combination of cues in some lists

Table 5.2 Estimations of β from the saturated Generalised Estimating Equation model for Study One (with robust standard error noted in parentheses) and corresponding odds ratios, with 95% confidence intervals. Design factors are noted in bold typeface.

Predictor		β(SE)	95% Confidence Interval Lower	Odds Ratio	Upper
Intercept		-1.00(0.54)			
SFX	***	0.40(0.59)	0.38	1.49	2.65
PREDICTABILITY	***	1.39(0.54)	2.95	4.00	5.05
ORDER	***	0.10(0.20)	0.72	1.11	1.51
SFX * LIST	***	-0.09(0.20)	0.52	0.92	1.32
ORDER * LIST	*	-0.01(0.06)	0.88	0.99	1.11
SFX * PREDICTABILITY	***	0.75(0.74)	-0.73	0.73	2.21
EXPERIENCE * **PREDICTABILITY**	*	1.58(0.54)	3.77	4.84	5.90
SFX * EXPERIENCE * LIST	*	-0.18(0.18)	0.48	0.84	1.19
SFX * PREDICTABILITY * LIST	***	0.38(0.28)	0.92	1.46	2.01
SFX * ORDER * EXPERIENCE * **PREDICTABILITY**	*	0.66(0.19)	1.56	1.93	2.30
SFX * ORDER * **PREDICTABILITY** * LIST	*	0.07(0.05)	0.96	1.07	1.19
SFX * EXPERIENCE * **PREDICTABILITY** * LIST	*	-0.09(0.10)	0.70	0.91	1.12

$*p < .05$, $**p < .01$, $***p < .001$

were easier than others. This is a commonly observed effect as the balancing of the original sentence list could not account for the effect of introduced experimental factors.

5.2.3 Objective results

An objective intelligibility metric was used to investigate the effect that the inclusion of the redundant non-speech audio objects may have had on the signal-level intelligibility. The selected objective metric was the GP [65], with an audibility criterion of 3dB (defined as how much louder the target needs to be compared with the masker in each spectro-temporal segment to be considered perceptible). The GP was calculated over the keyword to confirm that the improvements in intelligibility when non-speech sounds were present were the result of priming effects of the SFX and that the introduction of the SFX to the babble mixture did not reduce the masking potential of the babble over the keyword. This was averaged across the 50 trials for each condition and is shown in Table 5.3. The average values for all conditions were very close and statistical analysis indicated no significant differences

between the conditions $[p > 0.05]$. This gives weight to the hypothesis that the improvements in word recognition rate are due to the priming cues themselves.

Table 5.3 Glimpse Proportion (GP) for each experimental condition in Study One, evaluated separately for the keyword and preceding speech portion of the stimuli (with standard error noted in parentheses).

	LP	HP	LP+SFX	HP+SFX
Keyword	13.56% (0.76)	12.53% (0.82)	14.23% (0.84)	12.42% (0.83)
Preceding	18.86% (0.61)	18.57% (0.53)	9.85% (0.44)	10.07% (0.53)

Given the significant interaction between SFX and PREDICTABILITY in the subjective results, the possibility that the overlapping of the SFX had degraded signal-level intelligibility was investigated. The GP for the preceding speech was also calculated which can be seen in Table 5.3. For the conditions with semantic cues only (conditions LP and HP) the GP is very similar. The GP for the conditions LP+SFX and HP+SFX were also very similar to each other however they were almost half that of the semantic only conditions, a reduction of 47.0% on average. A Kruskal-Wallis test was used to determine whether the GP scores were significantly different for each experimental condition (as GP scores were found not to be normally distributed in some experimental conditions). There was no significant difference between the keyword scores with Kruskal-Wallis Chi Squared statistic = 3.7. For the preceding speech scores, a significant difference was found, with a Kruskal-Wallis Chi Squared statistic = 121.6, [p < 0.001]. Post-hoc analysis was then performed using the Dunn test. This showed that conditions where SFX were present significantly differed to those without $[p < 0.001]$, whilst there was no significant difference in GP due to predictability (i.e. HP and LP scores were not significantly different). This result was corroborated using another objective intelligibility measure, the Speech Intelligibility Index [290]. This metric showed on average $SII = 0.35$ for conditions with semantic cues only and $SII = 0.24$ for those with SFX. This represents a relative degradation in objective intelligibility between the conditions with and without redundant non-speech sounds of 31.6%. As such it is likely that the low odds ratio was due, at least in part, to the increased energetic masking caused by SFX overlapping the preceding speech effectively reducing the SBR.

The amount that the speech would need to be increased to compensate for this effective SBR reduction was investigated. By increasing the SBR and calculating the respective GP, the required increase in speech level required to meet $GP = 18.7\%$ (as in conditions LP+SFX and HP+SFX) was found to be $SBR = +2.5$dB. This represents a 4.5dB increase. This value is shown along with the GP for the experimental $SBR = -2$dB in Figure 5.4. To determine whether this value is likely to translate to other SBRs, the relationship between SBR and

GP was modelled in the range −15dB to +40dB (found using non-linear regression with a sigmoid fit). This is also shown in Figure 5.4, in the range −15dB to +15dB. Figure 5.4 indicates that a 4.5dB increase is likely to compensate for effective SBR reduction at higher SBR (until ceiling effects are reached). As the SBR decreases from the −2dB experimental SBR, the floor conditions are reached indicating the 4.5dB value may not hold in this region.

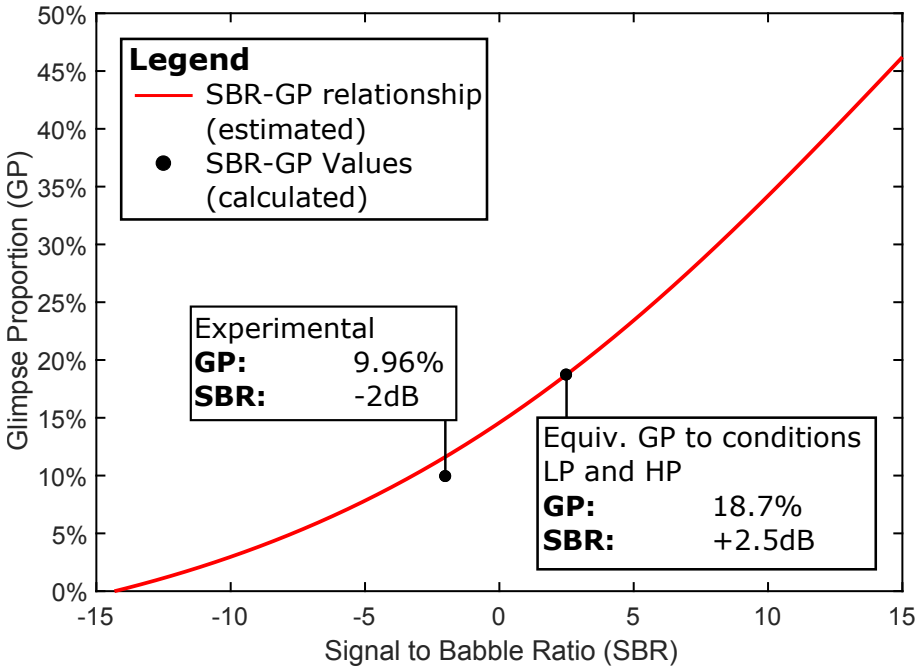


Figure 5.4 Equivalent speech to babble ratio for condition High Predictability with non-speech obejcts conditions (HP-SFX) required to achieved $GP = 18.7\%$ for the preceding speech, shown with relationship between Glimpse Proportion (GP) and speech to background ratio (SBR).

There are other plausible explanations for the interaction between SFX and PREDICTABILITY. As these are the dominant effects it is possible that participants were experiencing ceiling effects even without one cue impairing the function of another. This seems unlikely however as the average WRR for the condition + SFX was only 73.7%. It is also possible that there is a cognitive component of the effects with cognitive load increased due to the extra energy expended switching attention between the two types of cues.

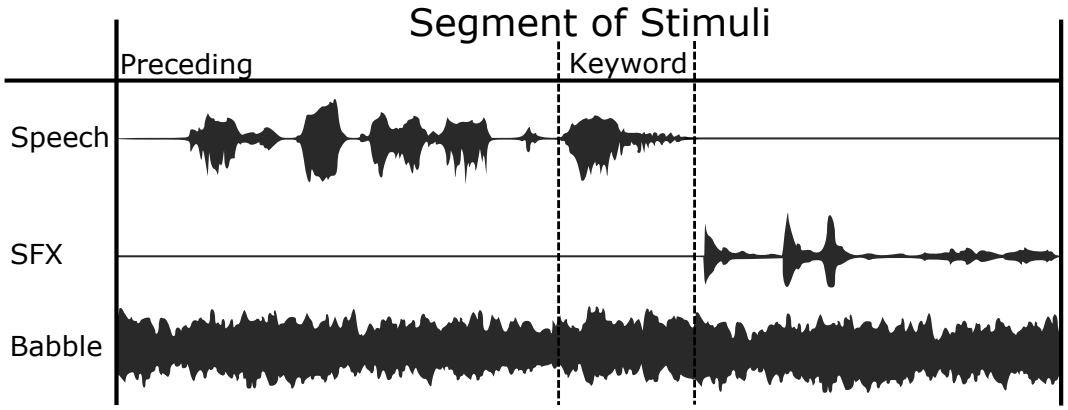


Figure 5.5 Example stimuli for Study Two, showing the alignment of the redundant, non-speech audio object (SFX) with references to the preceding speech and keyword. Note that the SFX does not overlap the speech and follows the keyword.

5.3 Study Two: Effect of separate, following SFX on keyword intelligibility

To further investigate what the independent effect of SFX on intelligibility may be, without the interaction of the speech, a second study was undertaken. This section describes the design, implementation, and results from this experiment.

5.3.1 Design of stimuli

This experiment was designed with the aim of maintaining as much in common with the experimental design from Study One as possible in order to facilitate the combination of the resulting data for comparison. The same sentence lists used in Study One in Section 5.2 and found in Appendix C, were used as well as the same SFX.

The key difference between the two studies was the location of the SFX relative to the keyword. Rather than overlapping the speech temporally as it does in Figure 5.1, it was placed after the keyword as can be seen in Figure 5.5. The babble was again normalised to -23dB^{LKFS} and set at an SBR of -2 dB. To ensure that the SFX were at the same level as in the first experiment, they were added to the babble and normalised to -23dB^{LKFS} (effectively ducking the level of the babble over the SFX).

Objective intelligibility analysis

For this experiment objective analysis was completed as part of the experimental design to ensure all conditions presented equivalent energetic maskers. This allowed for the effect of

SFX alone on intelligibility to be isolated. As in Section 5.2.3 this was performed separately over the keyword and preceding speech. The results of this can be seen in Table 5.4.

There was no change in the GP scores for condition HP and LP. It can be seen from Table 5.4 that the scores for the preceding speech in condition HP+SFX and LP+SFX are now within one standard error of conditions LP and HP. For both keyword and preceding speech, there was no significant difference in the GP scores between each of the experimental conditions, evaluated using Kruskal-Wallis test yielding a chi-squared = 0.3 and 5.1 for keyword and preceding speech respectively.

Table 5.4 Glimpse Proportion (GP) for each experimental condition in Study Two evaluated separately for the keyword and preceding speech portion of the stimuli, with standard error noted in parentheses.

	LP	HP	LP+SFX	HP+SFX
Keyword	13.56% (0.73)	12.50% (0.83)	14.92% (0.83)	12.69% (0.77)
Preceding	18.86% (0.61)	18.57% (0.53)	18.39% (0.53)	18.31% (0.56)

5.3.2 Implementation

The experiment was implemented in the same manner described in Section 5.2.1, including use of the same listening room and speaker set-up seen in Figure 5.2.

5.3.3 Results

Cohort

The cohort was comprised of 24 native English speakers (6 females and 18 males) who had self-reported normal hearing participated; 11 participants were aged $18-29$ yrs, 8 were aged $30-39$ yrs and 5 were aged 40 or older.

The same definitions of listening experience used in Study one (Section 5.2.1) were used here. Participants again had a variety of listening test experience with 10 being naive listeners, 10 being experienced listeners and 4 having moderate listening test experience. 10 of the participants also participated in the first part of the study (either the pilot or the main experiment), however a minimum of 9 months had elapsed between the participant partaking in each study.

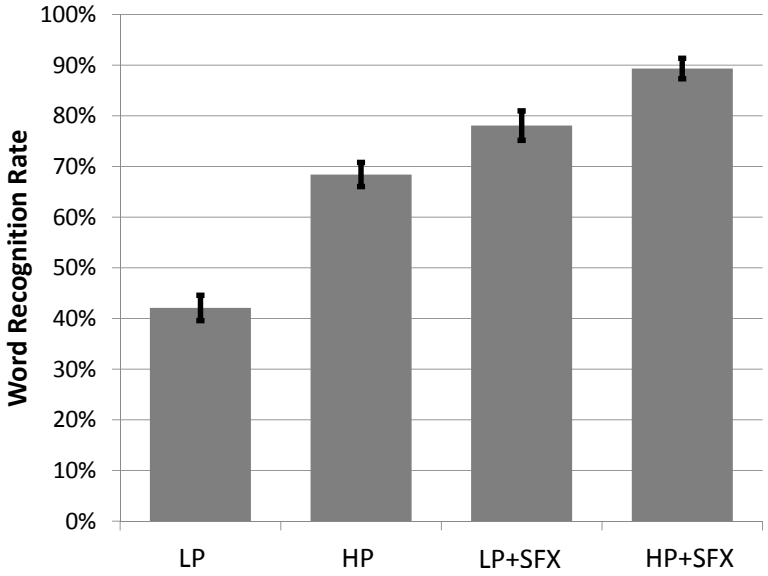


Figure 5.6 Mean word recognition rate ($n = 24$) for each experimental condition in Study Two shown with standard error bars and where LP indicates low predictability sentences, HP indicates high sentences and +SFX indicates the inclusion of relevant, non-speech sounds.

Subjective results

Figure 5.6 shows the mean WRR with standard error bars. Condition LP again has the lowest mean WRR of 42.1%. Condition HP gives a 62.6% improvement in recognition relative to condition LP, increasing the WRR to 68.4%. For condition LP+SFX the WRR increases to 78.1%, an improvement of 85.5% compared to the condition LP. This is a larger relative improvement than seen in the previous experiment and indicates the cues from the SFX may give greater benefit even when the preceding speech isn't meaningful. This suggests that some portion of the interaction seen in Study One may be attributable to the additional cognitive load incurred by switching attention between speech and the SFX. Condition HP+SFX shows a WRR of 89.3%, an improvement of 112.3% from the control condition, and a 14.4% and 30.5% improvement from the conditions with only non-speech and semantic cues respectively. All conditions had higher WRR than in the first experiment. This means that, from comparisons between the raw word recognition rates, the effect of not overlapping the SFX with the speech is not immediately apparent. However it does indicate that ceiling conditions were likely not reached in condition HP+SFX in Study One.

Statistical analysis of subjective results

As in the previous experiment, significance and effect size of the design and additional factors were evaluated utilising a saturated GEE model, the results of which can be seen

Table 5.5 Estimations of β from the saturated Generalised Estimating Equations model for Study Two (with robust standard error noted in parentheses) and corresponding odds ratios, with 95% confidence intervals. Design factors are noted in bold typeface.

Predictor		β(SE)	95% Confidence Interval Lower	Odds Ratio	Upper
Intercept		-1.25(0.55)			
SFX	***	1.6 (0.80)	3.42	4.99	6.55
PREDICTABILITY	***	2.37(0.73)	9.29	10.70	12.18
LIST	***	0.45(0.19)	1.20	1.58	1.96
EXPERIENCE	*	0.34(0.50)	0.41	1.40	2.38
SFX * LIST	**	-0.09(0.37)	0.19	0.92	1.64
SFX * BOTH	*	-0.19(5.06)	-9.09	0.83	10.75
PREDICTABILITY * LIST	***	-0.70(0.24)	0.03	0.50	0.96
SFX * ORDER * EXPERIENCE	*	0.25(0.28)	0.74	1.29	1.84
SFX * PREDICTABILITY * EXPERIENCE	*	0.02(0.73)	-0.43	1.02	2.46
SFX * PREDICTABILITY * LIST	**	-0.06(0.32)	0.31	0.95	1.58
* **PREDICTABILITY** * LIST * BOTH	*	-0.838(0.53)	-0.60	0.43	1.47
EXPERIENCE					
ORDER * LIST * BOTH * EXPERIENCE	***	0.04(0.15)	0.74	1.04	1.35
SFX * PREDICTABILITY * ORDER * BOTH * EXPERIENCE	**	1.30(0.91)	1.89	3.67	5.40

$^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$

in Table 5.5; again only the significant factors are reported. Given that some participants had taken part in both the first and second experiment, the additional factor BOTH was included in the model. The design factors are both highly significant, with the odds ratio for the non-speech cues again being approximately half the magnitude of the semantic cues. A notable omission from Table 5.5 is the interaction between SFX and PREDICTABILITY which is no longer significant indicating that when the speech and SFX do not overlap, their effects become additive (unlike in Study One). LIST, which was not a significant effect in the first experiment, was significant with the second largest odds ratio outside of the design factors, of 1.58. It indicates that participants found the first list of sentences harder and the sixth list easier, regardless of the presentation order. It is possible that without the effects of the interaction between cues, this effect has become more prominent or may be a random variation due to differences in the participant population. EXPERIENCE with listening tests

was seen to be a significant factor again, though most of the interactions with EXPERIENCE were only weakly significant and had odds ratios close to one. or not participants had taken part in the first experiment was not a significant predictor on its own however the interaction between SFX and BOTH was significant. This indicates that some of the participants likely remembered the SFX from the previous experiment. However given the exceptionally large confidence interval on this value, it is apparent that this was not true of all participants who had taken part in both experiments.

5.4 Comparison of Study One and Two

To evaluate how overlapping the SFX and speech affects the intelligibility to develop a model for this effect, the data-sets were partitioned to only include the SFX conditions, HP+SFX and LP+SFX. The data for Study One and Two were then concatenated. After performing checks to ensure complete separation and multicollinearity were not present, a saturated model was built. In this model only design factors from Table 5.1 and the additional variable EXPERIENCE were included; all others were treated as random effects (being LIST, ORDER and BOTH). This was done to create a generalisable model. These factors would form part of the random variation in the population rather than controlled factors or factors which could be easily estimated. Furthermore, as seen in the previous sections, the majority of these factors only have a small effect on the final results.

Table 5.6 Estimations of β from the saturated Generalised Estimating Equations model for Study One and Two (with robust standard error noted in parentheses) and corresponding odds ratios, with 95% confidence intervals. Design factors are noted in bold typeface.

Predictor		β(SE)	95% Confidence Interval		
			Lower	Odds Ratio	Upper
Intercept		1.05(0.10)			
PREDICTABILITY	***	1.00(0.12)	2.47	2.71	2.95
EXPERIENCE	***	0.23(0.06)	1.16	1.26	1.37
OVERLAP	***	-0.95(0.11)	0.15	0.39	0.62
PREDICTABILITY * OVERLAP	*	-0.36(0.15)	0.41	0.70	0.99

$^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$

To determine the optimal model a stepwise forward algorithm was utilised, using the R function `dredge` from the package MuMIn [326]. Given that GEE are not based on likelihood ratios, to evaluate the model fit at each step the Quasi-Likelihood Criterion (QIC)

was used rather than a more standard likelihood ratio. The best fitting model determined by the algorithm had a QIC = 5087 and significant factors which can be seen in Table 5.6.

5.4.1 Creating a generalisable model

From Table 5.6 we can use the estimated β values to get the regression model seen in Eqn. 5.2.

$$\begin{aligned} \text{logit}(\mu) &= 1.05 \\ &\quad +(1.00 \cdot \textsc{Predictability}) \\ &\quad +(0.23 \cdot \textsc{Experience}) \\ &\quad -(0.95 \cdot \textsc{Overlap}) \\ &\quad -(0.36 \cdot \textsc{Overlap} \cdot \textsc{Predictability}) \end{aligned} \tag{5.2}$$

From Eqn. 5.2 and Table 5.6, it can be seen that not only does OVERLAP interact with PREDICTABILITY, but it also has an effect on its own. As the SFX only overlaps the preceding speech, not the keyword, recognition of the LP keyword should not be affected if the only effect of the overlapping SFX has is energetic masking. This result suggests that when SFX overlaps speech, even if the speech is low predictability and does not semantically aid keyword recognition, it has an effect on the likelihood of correctly identifying target words. This may be because the temporal overlap of SFX and speech not only introduces energetic masking (as seen in Section 5.2.3) but also has cognitive effects – informational masking or increases in cognitive load due to switching attention.

Noting that Experience has three levels whilst the other factors only have two, we can see that EXPERIENCE = 1 (moderate experience) gives a small benefit. EXPERIENCE = 2 (experienced listener) yields double that benefit. Both these effects are smaller than the effects of PREDICTABILITY and OVERLAP.

To investigate the effect of utilising only the design factors and the additional factor EXPERIENCE to create the model, the predicted probabilities were compared with the empirical probabilities. The results of this can be seen in Figure 5.7. It can be seen that for naive and experienced listeners, the empirical and modelled probabilities agree very closely, with the empirical value falling inside the 95% confidence interval for all the modelled values. The values for the moderate experience differ somewhat, with the empirical probabilities outside the modelled 95% confidence interval for the No Overlap - HP+SFX condition. This

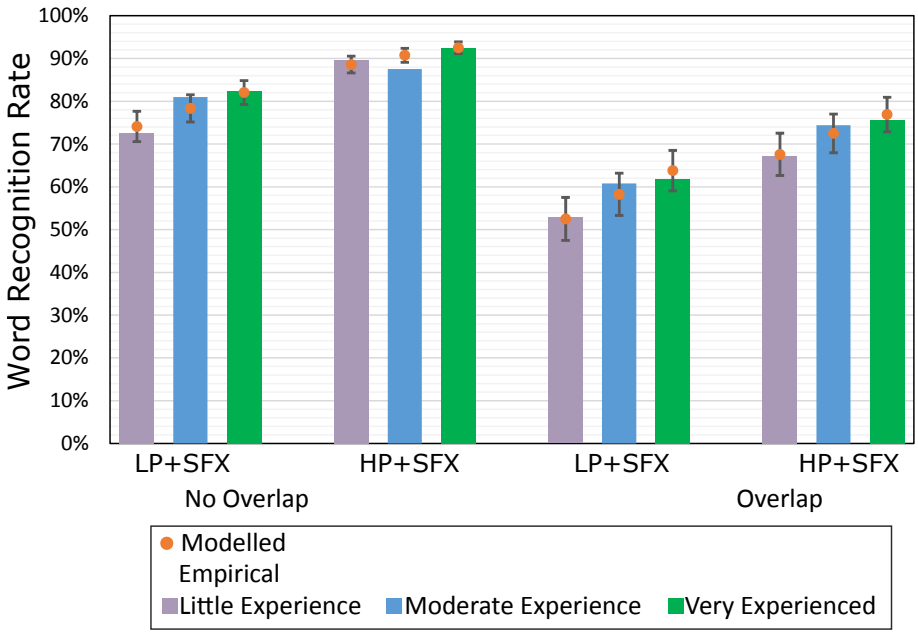


Figure 5.7 Empirical probabilities for correct keyword identification in conditions HP+SFX and LP+SFX shown for both Overlap and No Overlap states and for listeners of varying experience. Modelled probabilities, predicted using Eqn. 5.2, and their 95% confidence interval are also shown.

is likely due to the fact that in Study Two there were only four participants with moderate experience and there was insufficient data to model that condition effectively.

A further factor which should be considered when interpreting these results is the effects of forward and backward masking, as Study One utilised a preceding SFX whilst in Study Two the SFX follows the keyword. Backwards masking, the effect of masking prior to a strong sound, tends only to be prominent for 20ms after the target speech [327] and, as such, has likely not majorly influenced the results. Forward masking, the effect of masking after a strong sound lasts, can have an effect for up to 200ms [327] and may have contributed to the energetic masking and other cognitive effects seen in the Overlap conditions. Additionally, the effect of combining the speech and SFX together into a coherent cognitive auditory object changes if the SFX precedes or follows the speech is undetermined.

5.5 Chapter summary

In this chapter, two studies were described which aim to quantify the effect redundant non-speech sounds have on the intelligibility of speech. These studies awee closely related but differ in the placement the non-speech, redundant audio object in the time domain: preceding

the target word and overlapping preceding speech (Study One) or after the keyword (Study Two). Using the stimuli and the Glimpse Proportion metric, an investigation of energetic masking was undertaken. Using Generalised Estimating Equations, a model for the effect of Listener Experience, Predictability and Overlap on speech with a redundant non-speech sound present was built. From these results we can conclude the following key points:

- The inclusion of a redundant, non-speech sound aids intelligibility even if the sound overlaps preceding speech.
- Overlapping useful redundant, non-speech sounds with speech preceding the target word creates a reduction in objective signal-level intelligibility equivalent to reducing the speech to background ratio by 4.5dB.
- Positive effects of redundant, non-speech sounds and semantic cues on intelligibility effect speech recognition almost independently when the non-speech sound does not overlap the speech.
- Inclusion of non-speech sounds, though beneficial, also appear to energetically and informationally mask speech.
- The modelling of a relationship between listening experience, predictability of speech and whether a redundant, non-speech sound overlapping the speech suggests that listening study participation has a positive effect on ability to use complementary intelligibility cues.

Chapter 6

Evaluating the relationship between redundant non-speech audio objects and broadcast speech intelligibility for a hard of hearing cohort

6.1 Introduction

This chapter conducts the third study in Part II investigating the research question from Section 3.6.1:

What is the relationship between redundant non-speech audio objects and broadcast speech intelligibility, for normal and hard of hearing listeners?

Chapter 5 concluded that the inclusion of redundant non-speech sounds can aid intelligibility in cohorts of normal hearing individuals. This chapter builds on this, using the mixed methodological approach developed in Chapter 4, to determine whether this effect is present in hard of hearing cohorts. It first makes slight modifications to the experimental methodology, to accommodate the needs of a hard of hearing cohort. Results from this study are then presented and indicate that an individual's audiometric thresholds in their better hearing ear[1] can predict whether redundant non-speech sounds aid intelligibility or are detrimental to it. A number of limitations to this experimental approach are identified and discussed. These form the basis of further methodological development in Chapter 7, which is then implemented in Chapter 8.

[1] The term 'better hearing ear' as used in this work refers the whichever ear has a lower pure-tone audiometric threshold

6.2 Study Three: Effect of overlapping, preceding SFX on keyword intelligibility for hard of hearing listeners

This section describes the initial implementation of the methodology developed in Chapter 4. First, the calibration procedure developed for hard of hearing listeners is described, as well as with the experimental set up. The listening test results are then outlined as well as the objective intelligibility metric values for each of the calibrated speech to background ratios (SBRs). The conclusions from this initial study and its limitations are then discussed.

6.2.1 Implementation

This study utilises the methodology used in Chapter 5, developed in Chapter 4. However, unlike for the normal hearing cohort in Chapter 5, the SBR was set for each participant to ensure that, given their hearing loss, the test was neither too easy nor too difficult. Calibration of the appropriate SBR was achieved by using a list of unused sentences without sound effects (SFX) from the normal hearing implementation. The calibration list started at the -2dB speech to background used in Chapter 5 with a normal hearing cohort and was altered in 1dB increments until the participant indicated that they could understand approximately half of the sentences presented to them. Once selected, this signal to background ratio was maintained throughout the test. The cohort selected a wide range of SBRs from -2dB up to +12dB. Participants were also allowed to make small modifications to the overall reproduction level. The alterations selected ranged between +4dBSPL and -2dBSPL from the original 69dBSPL level used in Chapter 5.

In addition to the main part of the experiment, an audiogram was included in the experimental method. This was used to characterise participants' hearing and act as an explanatory variable in the analysis. The pure tone audiogram determines the lowest sound pressure an individual can perceive, termed the absolute threshold, at a given set of frequencies. This is usually conducted over headphones and performed individually for each ear. These thresholds are often averaged over a number of frequencies to yield a pure tone average (PTA) which is commonly used to define the severity of an individual's hearing loss [124]. An audiogram was selected as it is the most commonly used tool to clinically characterise hearing loss and it utilises a standardised procedure.

Only half the sentences from the implementation in Section 5.2 were used (Lists 3 and 4, as seen in Appendix C). This was to ensure that the total length of the experiment, inclusive of the audiogram and calibration procedure, was less than one hour so as not to induce

listener fatigue. Participants who had been fitted with a hearing aid were encouraged to wear it during the test if they usually wore it whilst watching television.

6.2.2 Results

Cohort

Fourteen predominantly older native English speakers took part. Audiometric thresholds for the frequencies 0.25Hz, 0.5Hz, 1kHz, 2kHz, 4kHz, and 8kHz were obtained for all participants using a Kamplex r27a Diagnostic Audiometer. The mean PTA, at speech frequencies (0.5-4kHz), across the cohort was 36dBHL (standard deviation = 21dBHL) and 49dBHL (standard deviation = 27dBHL) for their better and worse hearing ears respectively. The cohort had significant variation in their hearing impairments, ranging from normal hearing thresholds with tinnitus or Ménière's disease to severe loss (using the definitions in Table 2.1 from [35]). The majority of the cohort had symmetric hearing loss (12 out of 14).

Perceptual results

As each participants' stimuli had a different SBR, absolute word recognition rates (WRR) as used in Chapter 5 would not allow for comparison between the participants. Instead the improvement in WRR was calculated, relative to the low predictability with no SFX (base) condition for each participant (seen in Eqn 6.1). This improvement was averaged over all participants and the mean improvement in keyword recognition for the high predictability sentences was 91.8%. There was a large variation in this value, having a standard deviation of 63.0%. The benefit was positive for all listeners except one for whom the high predictability sentences made no difference. This benefit compares closely with previously reported results for hard of hearing listeners where high predictability sentences increased WRR from 28% to 70% (at 80dBHL and -1dB SBR) [77].

$$\text{Improvement} = \frac{\text{WRR target condition} - \text{WRR base condition}}{\text{WRR base condition}}\% \tag{6.1}$$

The mean improvement when SFX were added to the low predictability sentences was 9.9%, much smaller than for normal hearing listeners who exhibited a mean improvement of 69.5%. There was also a large amount of variation in this result with a standard deviation of 42.4%. For some participants, the SFX either degraded word recognition rate or had no effect. The addition of SFX to the high predictability sentences also offered only a small mean improvement of 13.18%. However, this had a smaller standard deviation of 19.0%

Table 6.1 Spearman two-tailed rank correlation between PTA (PTA, 0.5-4kHz) in better and worse ears, SBR and improvement in WRR when SFX are included for low and high predictability sentences.

	Better Ear PTA	Worse Ear PTA	Speech to Background Ratio (SBR)	SFX Improvement: Low Predictability
Speech to Background Ratio (SBR)	0.647*	0.629*	—	—
SFX Improvement:				
Low Predictability	-0.857***	-0.709**	-0.707**	—
High Predictability	-0.045	0.057	-0.103	-0.045

$^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$

and was of a similar magnitude to the improvement exhibited by normal hearing listeners of 18.7%.

Correlation analysis between the experimental factors was performed and is seen in Table 6.1. The aim of this analysis was twofold. First, to investigate whether the selected SBR was related to the participants' PTAs. Secondly, to determine whether the degree to which SFX were beneficial could be explained by how audible the SFX was (given the selected SBR and the participants' degrees of hearing loss). Normality of the variables was first assessed using the Anderson-Darling test for normality. As some of the variables did not meet the normality criterion, Spearman's rank correlation coefficient was used to evaluate the relationship between the different variables.

Table 6.1 indicates that the selected SBR is dependent on the PTA in both the participant's better and worse hearing ears. The degree of improvement (or degradation) which the SFX had on word recognition rate for low predictability sentences is strongly correlated with the participants' better ear hearing. It is also correlated, though less strongly, with the worse hearing ear and selected SBR. Figure 6.1a) shows a scatterplot of the PTA in the participant's better hearing ear against the SFX improvement for low predictability sentences. It can be seen that there is a monotonically decreasing relationship between the SFX improvement and better ear hearing. To determine whether better ear hearing alone was a predictor for the benefit of SFX inclusion in low predictability sentences, partial correlation analysis was also performed and can be seen in Table 6.2. It can be seen that when the effects of the worse hearing ear and the SBR are controlled for, the participant's PTA in their better hearing ear remains a predictor for how beneficial SFX are to word recognition rate in low predictability speech.

Table 6.2 Partial correlation between SFX improvement for low predictability sentences and pure tone averages (PTA, 0.5, 1, 2 &4kHz) in better and worse ears and SBR, using Spearman's two-tailed rank coefficient.

	Better Ear PTA	Worse Ear PTA	Speech to Background Ratio (SBR)
SFX Improvement: Low Predictability	-0.671*	-0.252	-0.304

$^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$

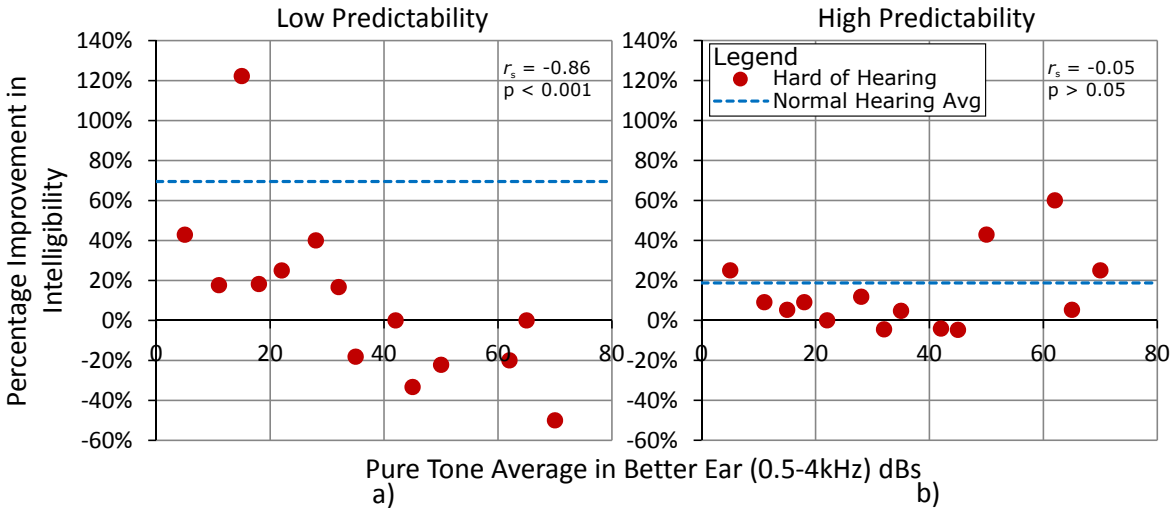


Figure 6.1 Scatterplot of pure tone average (PTA, 0.5, 1, 2 &4kHz) in their better hearing ear against improvement in word recognition rate when SFX were included for a) low predictability sentences and b) high predictability sentences. Average improvement demonstrated by normal hearing listeners in Study 1 from Chapter 5 and Spearman's rank correlation coefficient are also shown.

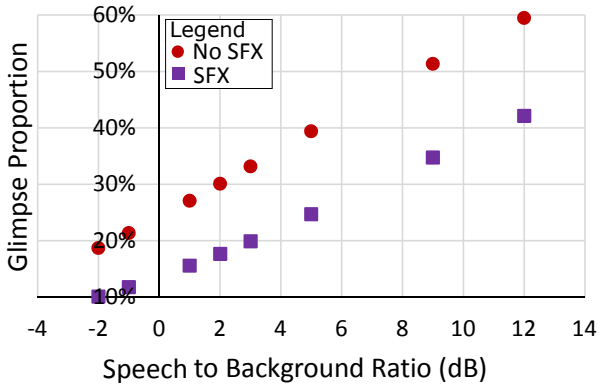


Figure 6.2 Mean glimpse proportion (GP) of the preceding speech for conditions with and without SFX, for each selected speech to background ratio (SBR).

It can be seen from Table 6.1 that unlike for low predictability speech, the degree to which SFX interact with high predictability speech does not correlate with hearing loss severity. Figure 6.1b) shows a scatterplot of better ear hearing and SFX improvement for high predictability sentences. Unlike the low predictability sentences there is no clear monotonic relationship across the range of hearing abilities. A monotonically decreasing relationship, similar to the one present for low predictability speech, does appear to exist in the region where the participants' PTAs are less than 50dBHL. Correlation analysis on participants with PTAs below 50dBHL was performed to determine the size and significance of this. This showed a significant montonically decreasing relationship, similar to the relationship seen for low predictability speech, with $[r_s = -0.784,\ p < 0.05]$.

Objective intelligibility analysis

The glimpse proportion (GP) for the keyword and the preceding speech were calculated separately as in Chapter 5. A three-way ANOVA between the two experimental conditions: predictability and presence of SFX, and the SBR was performed for both the keyword and preceding speech. This allowed for the effect of the SBR to be partialled out as changes in this produced the most significant differences between GP scores. For the keyword, conditions with SFX exhibited a slightly higher mean GP, though this difference was only weakly significant $[F = 5.5, p < 0.05]$. For the preceding speech, Figure 6.2 shows the range of GP values at each SBR for conditions with and without SFX. It can be seen that at all SBRs there is a large, and strongly significant, difference between the glimpse proportions when SFX are present and absent $[F = 1468.3, p < 0.001]$. Having controlled for the effect of different SBRs, these results mirror those seen for the normal hearing stimuli.

6.2.3 Discussion

From these results, it can be seen that for low predictability speech, the participants' SBRs in their better hearing ear is the strongest predictor for how beneficial redundant non-speech sounds are. As the stimuli was reproduced monaurally with the babble, speech, and SFX co-located, it is reasonable that performance would be dominated by the participant's better hearing ear. This relationship appears to approach the level of benefit exhibited by normal hearing listeners, as PTAs approach 0dBHL.

Interestingly, for some participants with higher PTAs, the presence of SFX did not improve but rather degraded word recognition rate below that of the control condition. Given that the level of the SFX was tied to the level of the babble, for participants who selected higher SBRs (predominantly those with higher PTAs), the SFX were presented at a

correspondingly lower level. The reasons that the SFX actively degraded intelligibility may be linked to this reduced audibility, as more of the listener's attention was required to identify the quieter sound and subsequently make use of it. Furthermore, given that the preceding speech did not relate to the SFX, the process of switching attention between the speech and the SFX may have resulted in increased cognitive load [328]. This increased load potentially impaired parsing of the speech and SFX compared with when the cognitive resources are mostly dedicated to the speech alone (in the babble only conditions). A similar hypothesis was proposed in [196] where the addition of music and SFX was shown to reduce knowledge transfer in multimedia content. Whilst [196] only studied normal hearing listeners, it is possible that this effect is more prominent in those with higher degrees of hearing loss. However, given that on average the keywords of conditions with SFX had slightly lower glimpse proportions and subsequently slightly more energetic masking, it is also possible that this was having a greater impact on those with higher PTAs. This may have contributed to the degradation in intelligibility for these participants.

For high predictability speech, it appears that for hard of hearing listeners with a PTA below 50dBHL, better ear hearing remains a useful predictor of SFX benefit. As with low predictability speech, it appears as PTAs approach 0dBHL, this relationship approaches the level exhibited by normal hearing listeners. However given the small size of the cohort and the large variability in their hearing impairments, it is possible that this trend may not be generalisable. As with the low predictability stimuli, some hard of hearing listeners found the presence of SFX degraded intelligibility. In addition to the possible distraction effects of the SFX, for high predictability speech, this degradation in intelligibility may be due to the SFX energetically masking the clues from the preceding speech (as indicated by the significantly reduced GP when SFX were present). However, unlike for low predictability speech, for listeners with PTAs above 50dBHL the SFX did not degrade intelligibility. It is possible in this condition that the preceding high predictability speech aided the listener in identifying the SFX, rather than the other way around. The overall effect being that, despite listeners' difficulty in identifying the SFX the SFX still acted as redundant information for determining the keyword. It is further evident that a more complex relationship between better ear hearing and SFX utility exists for high predictability speech than for low predictability, which warrants further investigation (and is addressed in Chapter 8).

6.2.4 Limitations to the current methodology

One of the limitations of this approach for hard of hearing listeners is that an individually calibrated SBR is required. This approach makes results more difficult to compare between participants and introduces potential error if the correct calibration level is not found. Fur-

thermore, it has been shown that the fluctuating masker benefit, the amount that the available glimpses of speech above the masker aid speech intelligibility, are dependent on the SBR [329, 330].

A solution to these problems would be to use a speech reception threshold approach, similar to that used for the R-SPIN by Wilson et al [190]. Such an approach would remove the need to calibrate individual ratios and allow all participants to utilise the same stimuli. Furthermore, this will allow the determination of a 50% speech reception threshold under each experimental condition for each listener. This will aid comparison between participant's results and may also facilitate better comparison between these results and other speech in noise studies.

Another limitation of the current method was that the level of the SFX was tied to the level of the babble masker. This approach gives an insight into the effect of the SFX on intelligibility of legacy content where the level of all non-speech elements are likely to be reduced together. However, as the level of the SFX is free to be altered within object-based content, this is not the most ecologically valid approach for this work. In addition to a multiple speech to background level paradigm, the SFX level should be kept independent from the babble level in the revised methodology.

Furthermore, whilst the R-SPIN stimuli has proved effective for comparing the effect of non-speech objects on intelligibility, a number of factors reduce its ecological validity. The first of these is the quality of the recording. The original recording was made onto magnetic tape resulting in the speech having a limited bandwidth (no content >8kHz); furthermore there is the presence of high frequency tape hiss on the recordings. Given that the majority of broadcast audio is recorded at 44.1kHz or higher, this stimuli is not representative of the frequency range or quality of broadcast content.

Furthermore, the speaker has an American accent. Whilst a significant portion of broadcast content in the UK is made in the USA, a British accent is more likely to represent an ecologically valid stimuli for British cohorts.

Whilst audiometric thresholds proved useful explanatory measures in this experiment, it is well documented that they are limited in their ability to characterise hearing loss. For this reason, the subsequent methodology will endeavour to utilise more varied and representative measures of hearing loss characterisation.

Addressing these issues required re-recording the R-SPIN stimuli as well as making changes to the arrangement of the stimuli and other measurements included in the methodology. These changes are discussed in Chapter 7.

6.3 Chapter Summary

This chapter has begun to explore whether the effect of redundant non-speech objects have the same effect on intelligibility for hard of hearing cohorts as exists for normal hearing listeners. The implementation detail of the methodology, developed in Chapter 4, for the hard of hearing cohort was outlined. The results of Study three demonstrated:

- An individual's audiometric threshold in their better hearing ear predicts the benefit of redundant non-speech objects in perceiving low predictability speech.
- For some listeners the presence of redundant non-speech objects degrade intelligibility which may be due to the participant specific speech to background ratios utilised in this study.
- For those with audiometric threshold less than 50dBHL in their better hearing ear, audiometric threshold remains a predictor for non-speech object utility in recognising high predictability speech.

However it is evident that a much more complex relationship between hearing ability, speech perception, and redundant non-speech objects exists for high predictability speech. There are a number of limitations which the current methodology experiences for use with hard of hearing individuals which include: the quality of the stimuli, the single speech to background ratio (calibrated to the listener), the limited information about the participants hearing ability, and linking the level of the non-speech sounds to the level of the background sound. These limitations will be addressed through further methodological development in Chapter 7 and utilised with a hard of hearing cohort in Chapter 8.

Chapter 7

Evaluating intelligibility: Further methodological developments for hard of hearing cohorts

7.1 Introduction

A number of limitations in the experimental method developed in Chapter 4 were identified in Chapter 6 when first implemented with a hard of hearing cohort. In particular are the limitations caused by use of a single speech to background ratio, yielding a percentage word recognition score, rather than a multiple speech to background ratio which gives a speech reception threshold. The quality of the stimuli used also presented a key issue.

This chapter will describe the work completed to refine the experimental methodology for hard of hearing cohorts. It will first outline the process to improve the quality of the stimuli, through re-recording the Revised Speech Perception in Noise test. This new version, termed the Re-recorded Revised Speech Perception in Noise test or R^2SPIN, is then validated with a normal hearing cohort. This shows that the new stimuli are still effective at evaluating the effect of semantic cues on intelligibility.

The development of the new R^2SPIN into a multiple speech to background ratio paradigm, producing a speech reception threshold, is then described. This also presents a methodology for generating recognition performance data for speech perception tests exploiting computational intelligibility metrics. Finally, alternate methods of characterising the hearing ability of hard of hearing cohorts is then discussed.

7.2 Development of the methodology

The five key methodological problems identified in Chapter 6 were:

1. Limited characterisation of participants' hearing loss.
2. The quality of the R-SPIN recordings.
3. The accent of the speaker in the R-SPIN recordings.
4. The linked level of the non-speech objects to the level of the SFX.
5. The single speech to background ratio (SBR) yielding a word recognition rate.

The use of multiple measures which characterise different aspects of hearing loss can address the first of these issues. The first section of this chapter describes the three additional clinical and research methods which were chosen: the Quick Speech in Noise test (QuickSIN), the SSQ49 (as used in Chapter 3) and a measure of temporal fine structure [67].

The limitations around the quality and accent can be addressed by re-recording the stimuli with a British speaker and high quality recording equipment. The process for this is described in the second section. This stimuli were then validated to ensure that the main factor under test was not unduly influenced by any characteristics of the new recordings. The process of this and the results are described.

To address the single SBR, the new stimuli were re-arranged into a multiple SBR paradigm following a similar methodology to the one previously used for the R-SPIN [190]. This approach produces a speech reception threshold, the SBR at which 50% of speech can be understood. A computational intelligibility metric, the glimpse proportion, was used to evaluate the level of masking of each stimuli at each SBR. This analysis was then used to distribute the stimuli from easiest to most difficult to smooth the psychometric curve. This process, along with other modifications to the experimental design, are described in the final section of this chapter.

7.3 Characterisation of Hearing Impairment

Three additional measures were selected to cover the widest range of hearing characteristics as was practicable. Two validated and clinically used measures were selected: pure tone audiometry and the Quick Speech in Noise (QuickSIN) [331] test. Increasingly, speech perception in noise tests are used clinically to provide a more complete measure of an individual's ability to understand speech in everyday scenarios [332]. The QuickSIN was selected as it is a measure used in clinical settings and acts as an independent measure of speech in noise with which the R-SPIN measures can be compared. The audiogram was

maintained in the test battery to facilitate comparison with other works and the results of Chapter 6.

In addition to these clinical measures, a measure of temporal fine structure was included as well as the full form of the 'Speech, Spatial and Qualities of hearing' – the SSQ49 – survey. The full form is used here, rather than the reduced SSQ12 used in Chapter 3, to gain a more in-depth view of the participants hearing. The measure of temporal fine structure was selected as many recent studies suggest a link between higher sensitivity to temporal fine structure and speech in noise performance [333, 256, 334, 335]. Finally, individuals were asked to report their level of hearing loss and other salient details about their use of assistive hearing devices.

The two new measures which were introduced are described in the following section. A detailed summary of the SSQ49 can be found in Section 3.2.2.

7.3.1 QuickSIN

QuickSIN is a modified version of the Speech in Noise test developed by Killion et al. [331]. It utilises the IEEE sentences spoken by a female speaker presented in four-talker babble noise. Each list contains six sentences with five key words per sentence. The sentences are presented at pre-recorded signal-to-noise ratios (SNR) decreasing in 5-dB steps from 25 (very easy) to 0 (extremely difficult). These SNRs are designed to encompass the point of 50% recognition for normal to severely impaired people. From performance at each level, SNR loss can be calculated with a normative value of 2 dB SNR loss for normal hearing individuals.

7.3.2 Temporal Fine Structure

Sensitivity to temporal fine structure (TFS) refers to the ability to perceive the rapid oscillations close to the centre frequency of a complex sound such as speech. It has been demonstrated that hearing impaired and aged listeners have a reduced ability to utilise TFS information [67]. A measure of TFS was included in order to explore the relationship between TFS and television speech understanding.

There are a number of different ways in which sensitivity to TFS can be estimated including interaural phase descrimination (IPD), pitch perception and speech recognition [334]. The tool used in this study was the TFS-AF test developed by Füllgrabe and colleagues [336, 337] which presents the low frequency IPD task developed by Hopkins and Moore [338] in an adaptive paradigm. This test determines the highest frequency at which a fixed interaural phase difference of 180° can be perceived as different from a phase difference of 0°.

This was selected as it was designed to address problems encountered with the administration of other similar tests where some listeners, particularly older listeners, are unable to complete the tasks and the results are not easily comparable across participants [337–339]. The method uses a 2 option forced choice paradigm and an adaptive procedure with a 2-up, 1-down stepping rule which terminates after 8 reversals and takes the geometric mean of the last 6 reversals to estimate the 71% point on the psychometric curve. An example result can be seen in Appendix 1.

7.4 R^{2}SPIN - Re-recording the Revised Speech Perception in Noise test

To address the accent, gender, and quality issues with the original R-SPIN recording, re-recording the stimuli was required. This section describes the methodology used to ensure that these changes did not alter the validity of the original test.

7.4.1 Methodology

Sentence validity

In the original test, to ensure word familiarity, the words were selected from [340] which contain the 30,000 most frequently used words. Whilst some sentences do show their age (*My TV has a 12 inch screen*), given their simplicity they were still deemed to be familiar and interpretable to modern listeners. Furthermore, retaining all the same sentence text helps to maintain the phonetic balance described in the following section.

Speakers and audio recording

Two native British English speakers, one male and one female, with extensive experience in broadcast and radio were selected. Each speaker recited all 400 sentences. The speakers were instructed to use a neutral tone and pace.

The audio was recorded in a quiet room at BBC R&D, Cardiff[1] using a Neumann TLM 193 microphone at a distance of 0.25m from the speaker. They were recorded into a Sadie digital audio workststation at a sample rate of 48 kHz and bit depth of 32 bit and saved as uncompressed .wav files.

[1]These recordings were conducted by Catherine Robinson.

Quality verification and post-processing

A quality assessment was made of the recorded sentences, to ensure speech clarity and correct pronunciation as well as check for recording artefacts. This assessment was undertaken over headphones by five listeners trained in critical listening, including the author. They were instructed to identify any of the following problems: mispronunciations (compared with the target sentences), rushed or slow delivery, speaking too loudly or softly, editing and recording artefacts as well as any miscellaneous problems. One assessor conducted the validation without the target sentence list to ensure the effects of priming did not prevent error identification. This was performed for both the male and female sentences and at least two of the assessors were used for each iteration. All problematic sentences identified were re-recorded using the same equipment and conditions described in Section 7.4.1. For sentences with mispronunciations, specific error notes were fed back to the speakers.

Post-processing was performed on the sentences using Adobe Audition [341]. Silences between sentences were removed manually. The sentences were aligned with the keyword of the original speech, so that the babble noise for all keywords (old and new) were as identical as practically possible.

Pilot

A pilot study with 6 normal hearing listeners was undertaken to determine the appropriate SNR for Sections 7.4.3 and 7.4.4. An appropriate SNR is defined as one where the HP sentences do not saturate at the top of the psychometric curve and similarly the LP sentences do not saturate at the floor. This is achieved by having an overall word recognition rate of $\approx 50\%$. Initially -2dB SNR was used with 3 pilot participants as in previous work [342]. These were seen to have a mean word recognition rate of 74.5%, averaged across both HP and LP sentences. This was reduced to -4dB SNR, for a further three pilot participants, reducing the mean value to 56.3% across all sentences.

7.4.2 Phonetic analysis

Phonetic analysis was conducted to ensure that the change of accent did not significantly change the phonetic balance of the lists. Sentences were transcribed into both Standard American Pronunciation and Received Pronunciation. [2]

[2]The transcriptions were completed by Katherine M. Tucker, a performer with postgraduate training in IPA.

As in the original manual for the R-SPIN [343], each type of vowel and consonant was summed. The totals, along with the totals from the original R-SPIN can be seen in Tables 7.1 and 7.2 respectively.

Table 7.1 Vowels usage in R^2-SPIN by the female (F) and male (M) speakers compared with vowel usage in R-SPIN by the original (O) speaker, grouped by place of articulation.

	High Front	Mid Front	Low Front	Mid Centre	High Back	Mid Back	Low Back	Rhotic*	Dip-thongs	Total
F	41	20	35	5	12	24	11	0	52	200
M	41	20	35	5	12	24	11	0	52	200
O	41	20	29	0	12	19	10	19	50	200

** otherwise referred to as r-coloured vowels*

Table 7.2 Consonant usage in R^2-SPIN by the female (F) and male (M) speakers compared with consonant usage in R-SPIN by the original (O) speaker, grouped by manner of articulation.

	Plosive (UV)	Plosive (V)	Plosive (UV)	Plosive (V)	Semi-vowel	Nasal	Affricate	Total
F	138	71	110	12	93	63	17	504
M	138	71	99	23	93	63	17	504
O	138	71	99	23	93	63	17	504

UV denotes unvoiced consonants and V denotes voiced consonants

It can be seen from Tables 7.1 and 7.2 that the total number of phonemes in the original and re-recorded versions are the same. The analysis conducted here uses finer grain categories than the original analysis [343] based on place of articulation (for vowels) and includes affricates (for consonants). The main difference between the received pronunciation and the original standard American pronunciation is the distribution of vowel types. A distinct characteristic of the standard American accent is rhotic, or r-coloured, vowels. As received pronunciation does not have rhotic vowels, these are distributed to other vowel categories. There is no difference between the new and older male speakers in consonant type. The main difference between the new male and female speakers is the tendency of the female speaker to de-voice fricatives when they occur as plurals (\z\–> \s\). However, as this behaviour was consistent across the stimuli by the female speaker, this characteristic of her speech was maintained.

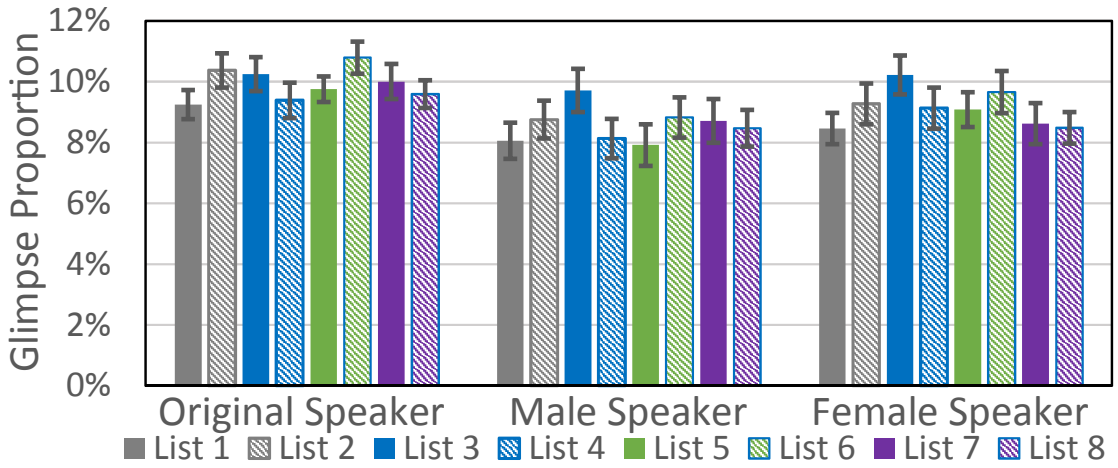


Figure 7.1 Glimpse proportion (GP) for each speaker: original in the Revised Speech Perception in Noise test (R-SPIN) and female and male speakers in the new re-recorded R-SPIN (R^2-SPIN), noted by list number with standard error bars shown.

7.4.3 Objective intelligibility analysis

Objective intelligibility metrics quantify signal and masker interactions and how they affect word level intelligibility (e.g. [344, 345]). This gives an initial insight into the possible intelligibility differences between lists without requiring large scale subjective evaluation. Both the original and new stimuli were analysed using an objective intelligibility metric called the glimpse proportion (GP) [65]. The GP quantifies the number of time-frequency regions of speech which survive energetic masking and reflects the local audibility of speech in noise.

Methodology

All sentences were normalised to -23dB^{LKFS} using the ITU-R BS.1770-2 specification [323]. This was also done for the multi-talker babble signal, allowing the -4dB SNR selected in the pilot (Section 7.4.1) to be set. This was done for both the original and new speakers, though only the new speakers were used for the subjective evaluation (Section 7.4.4). The GP was calculated in Matlab and calculated twice for each sentence: once over the keyword only and once over the preceding speech only. This allowed analysis of the effect of energetic masking on the keyword itself and the speech preceding the keyword (given its importance in the HP sentences). Keywords were aligned as described in Section 7.4.1 so that direct comparisons could be made; due to differences in accent and pacing, the remainder of the sentences were not aligned.

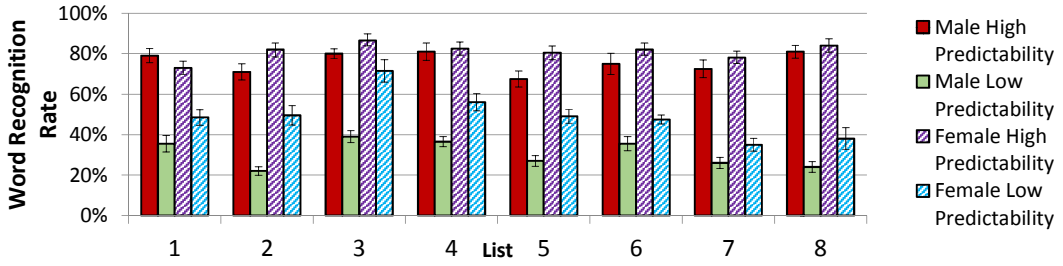


Figure 7.2 Mean word recognition rate for the female and male speakers in the re-recorded R-SPIN (R^2-SPIN), noted by list and sentence predictability, number with standard error bars shown.

Results

Table 7.1 indicates some inter-list difference between the mean keyword GP. The distribution of the means across the lists is reasonably constant (i.e. the lists with higher GP are roughly similar across the speakers). The GP results indicate which of the lists and speakers are likely to have higher energetic masking (i.e. lower GP), and as such, likely to be more difficult. Tables 7.3 and 7.4 show the mean GP for keyword and preceding speech for each speaker and list respectively. Table 7.3 shows that List 3 has the highest mean GP for the keyword, followed by List 6. The list with the lowest keyword GP is List 1. For the preceding speech the lists with the highest GP are 4 and 8, with List 6 being the lowest. Table 7.4 shows that the original speaker has the highest mean GP and easiest level of energetic masking. For the preceding speech, this order changes with the male speaker (most difficult for the keyword) having the highest GP for the preceding speech. The female speaker's preceding speech shows the lowest average GP. The GP of the keywords are smaller and more variable, as in previous studies [342] and this results from the keywords' short time window which makes it more vulnerable to the fluctuations in the masker.

A two-way ANOVA was performed using Matlab to determine whether these apparent differences were significant and was performed separately for the keyword and preceding GPs. A highly significant difference between speakers was shown for both keywords [$F = 9.99, p < 0.001$] and preceding speech [$F = 185, p < 0.001$]. Post-hoc testing showed that for the keywords the male and female speakers were significantly different to the original speaker but not to each other. For the preceding speech, all speakers were significantly different. The lists showed a weak significant difference for both the keywords [$F = 2.09, p < 0.05$] and the preceding speech [$F = 2.13, P < 0.05$]. Interaction effects were not significant.

Table 7.3 Mean Glimpse Proportions (GP) for R^2-SPIN (male and female) shown by list and decomposed into GP for the keyword and preceding speech.

List	Key Word	Preceding Speech
1	8.6%	15.3%
2	9.5%	15.4%
3	10.1%	15.6%
4	8.9%	16.0%
5	8.9%	15.9%
6	9.8%	15.2%
7	9.1%	15.7%
8	8.9%	16.2%

Table 7.4 Mean Glimpse Proportions for R^2-SPIN (male and female) shown by speaker, averaged over the lists.

	Key Word	Preceding Speech
Original	9.9%	15.3%
Male	8.6%	17.9%
Female	9.1%	13.8 %

7.4.4 Subjective intelligibility analysis

This section's small scale subjective analysis complements the objective analysis with a human normal hearing population (as validation of all possible experimental combinations with a large human population is impractical). The subjective analysis only evaluated the new speakers and used the same stimuli as prepared for Section 7.4.3.

Methodology

The sentence lists were presented to the participants in a pseudo-random order. Each participant received all eight lists, four with a male speaker and four with a female and each gender using a keyword only once. Lists with the same keyword were separated as much as practicable to reduce learning effects. The experiment was broken into four parts each containing two lists, one of each gender, and after each part, participants were offered breaks to avoid fatigue. The order in which the lists were presented was also pseudo-randomised with each of the sixteen possible orders being presented once.

Tests were undertaken in listening rooms at two locations: BBC R&D and the University of Salford. To ensure these different locations did not introduce room artifacts, the stimuli were presented to the listener over a set of Sennheiser HD 800 headphones (as opposed

to loudspeaker presentation as used in the previous sections of this work). The stimuli were presented with speech and noise co-located at 0° and reproduced at a level of 69dBA (calibrated using pink noise). Participants used pen and paper to record each keyword as in previous implementations in this work (Chapters 5 and 6).

Participants

Only participants with normal hearing (self-identified) and with English as their first language were recruited. Participants were naive listeners, defined here as having participated in less then 10 previous listening experiments. This was to reduce the possibility of learning effects from previous exposure to the stimuli. Only participants who were 35 years old or younger were recruited, to avoid the possibility of un-diagnosed age-related hearing loss. 5 females and 11 males meeting these criteria were recruited with a median age of 23 years old and mean age of 24.

Results

The word recognition rate results can be seen in Figure 7.2 and, for all except the female speaker on list three, the characteristic R-SPIN improvement of 30-40% between LP and HP sentences is maintained. Variation existed between the lists and speakers which is to be expected given Section 7.4.3. We see that, for all except the male HP sentences in list 1, the female speaker has a higher word recognition and this is particularly pronounced for the LP sentences. Given that the female speaker has a higher GP on average than the male speaker over the keywords – which are likely to dominate the result for LP sentences – this could be attributed to lower energetic masking. To investigate this effect further, the 8 female lists were repeated with three listeners at a more challenging -6dB SNR. This reduced the average value of the word recognition rate from 81.1% to 65.5% and 49.4% to 31.8% for HP and LP respectively. For list 3, which Figure 7.2 shows has a particularly high word recognition rate for the LP sentences, the word recognition rate reduces from an average of 71.5% to 52.0%.

As in Chapter 5, given the dichotomous outcome variable (right or wrong), a standard ANOVA could not be performed. Again, generalised estimating equations (GEE) were used here to analyse the results. The data was tested for multicollinearity and complete separation and was found not to violate these assumptions. In addition to the two design factors of the experiment (predictability and speaker gender), further predictors were investigated: LIST, ORDER OF PRESENTATION, GP (keyword and preceding speech), RETAKE (whether the sentence was the original re-recording or a retake made after Section 7.4.1 and PARTICIPANT AGE and GENDER. To ensure interpretable results given the large number of factors, a

Table 7.5 Odds ratio from the first order Generalised Estimating Equations (GEE) model for the re-recorded Revised Speech Perception in Noise test (R^2-SPIN) (with robust standard error noted in parentheses) and corresponding 95% confidence intervals.

Factor	Odds Ratio	95% Conf. Interval	
Retake	0.73	±1.05	**
List	0.96	±0.22	*
Speaker Gender	0.26	±0.73	***
Predictability	2.41	±1.10	***
Keyword GP	1.10	±0.10	***

$^*p < .05$, $^{**}p < .01$, $^{***}p < .001$

model containing only up to second order interactions was developed (using the package geepack in R [325]). Wald's test was then used to determine which factors offered significant improvements to the power of the model.

Table 7.5 shows significant first order factors and confirms that PREDICTABILITY, the factor under evaluation in R-SPIN, is significant and has the largest effect size (largest odds ratio). This confirms that the new stimuli would still be valid for evaluating the research questions in this work. SPEAKER GENDER and KEYWORD GP are also both highly significant with SPEAKER GENDER also having a large effect size. Given the results in Section 7.4.3, this was likely caused by the speaker based differences in energetic masking. DIFFERENCES BETWEEN LISTS was also significant, though only weakly and with a small effect size. RETAKES were also significantly different, though given the large confidence interval of the odds ratio, it is likely this is capturing the variation inherent in the speech rather than the variation due to re-recording. The interactions with high significance ([$P < 0.01$]) were: [LIST*PREDICTABILITY], [PARTICIPANT GENDER*AGE], [AGE*ORDER], [GP*ORDER], [GP*PREDICTABILITY], [LIST*KEYWORD GP] and [SPEAKER GENDER*KEYWORD GP]. These interactions had odds ratios in the range of 0.85 to 1.08, indicating their effects were small.

From these results we can conclude that the improvements in recording quality, speakers, and accent have not significantly affected the main factor under test: the sentence predictability. This is further reinforced by the maintenance in the new recordings of the characteristics 30-40% improvements in word recognition scores between high and low predictability sentences. Having resolved these issues with the stimuli, R^2-SPIN can now be developed into a multiple SBR paradigm.

In addition to the work completed here, the new stimuli and all resources were released under a Creative Commons license and can be obtained at: https://github.com/bbc/r2spin

7.5 Adaptions to the experimental implementation

This section describes the rationale for modifications to the experimental design, such as the static level of the SFX. It also details the process which was undertaken to construct the stimuli into a multiple SBR paradigm.

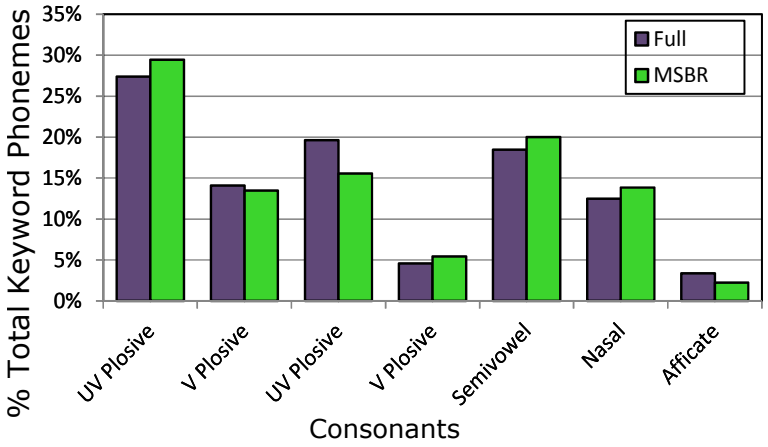
To ensure that the current results were comparable to the previous studies in this work, it was decided that the 12 speaker multi-talker babble would be used. Furthermore, given the large effect size and difference in performance between the speaker genders, it was decided that for the following study only the male speaker would be used. This also allowed the number of experimental factors to be reduced.

7.5.1 Selected sentence stimuli

It was shown in Chapter 6 that the relationship between non-speech audio objects and high predictability speech was much more complex than for low predictability speech. For this reason, the focus of the following study was on high predictability speech. Subsequently, three types of stimuli were chosen: high predictability sentences (HP), high predictability with non-speech objects (HP+SFX) and low predictability sentences (LP). The latter included to provide a baseline speech reception threshold for each participant. This gave a reference point for the benefit they gain from semantic and non-speech object-based context.

From the 400 re-recorded sentences, 36 high predictability sentences with keywords which could be easily linked to SFX were selected. An additional set of 36 LP and their corresponding HP sentences were also selected. The majority of these sentences were selected from Lists 1 and 2 to maintain as much of the original sentence balance as possible.

An analysis of whether the selected sentences were a representative phonetic subset of the full test was performed. The results of this can be seen in Figure 7.3. It can be seen that the percentage of each phoneme used in the multiple SBR subset closely matches the balance in the full form of the R^2SPIN (with the new male speaker). Attaining a perfectly representative subset was not possible due to limitations caused by selecting the most appropriate keywords for linking to SFX. The phonetic balance was not maintained at each individual SBR. The reason for this was that the allocation of keywords to different SBRs was determined by their GP, rather than by their phonetic content for reasons outlined in Section 7.5.3. The final sentence lists can be found in Appendix D.



UV and V denote unvoiced and voiced phonemes respectively

Figure 7.3 Percentage of total consonant content by phoneme group for the full and multiple speech to background ratio (MSBR) forms of the R^2SPIN.

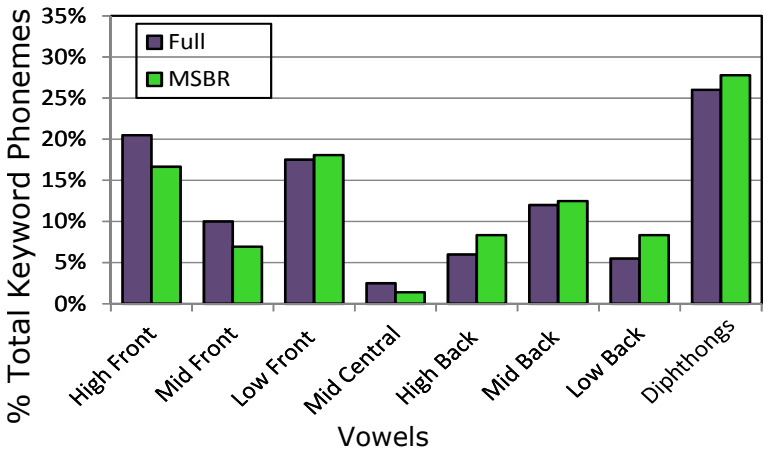


Figure 7.4 Percentage of vowel phonetic content by phoneme group for the full and multiple speech to background ratio (MSBR) forms of the R^2SPIN.

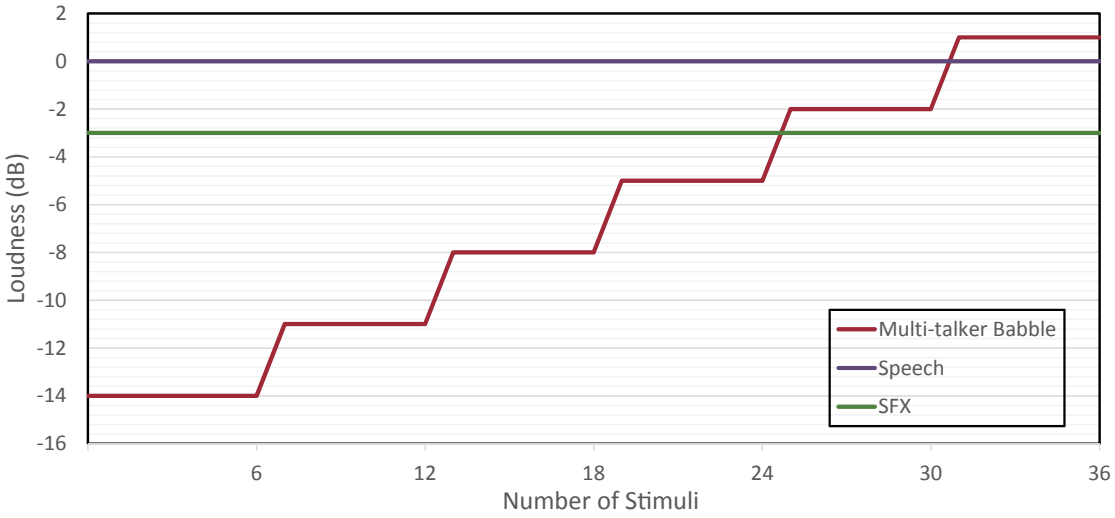


Figure 7.5 Visualisation of each level of the multiple speech to background ratio version of the R^{2}SPIN stimuli, showing the multi-talker babble, speech and redundant non-speech object (SFX) levels in dB.

7.5.2 Level independence of the SFX

One of the key problems identified in Chapter 6 was that by linking the SFX level to the level of the babble, as the babble was lowered the SFX may have become unrecognisable or simply inperceptible to the listener. To address this, the SFX was set at a static level (with reference to the speech) for all background ratios. This was selected to be -3dB, as informal consultations with production staff indicated that this represented an ecologically valid level for a salient SFX in television content; a visualisation of the relative levels of each audio component can be seen in Figure 7.5.

As it was necessary for the level of the SFX remain static, their loudness had to be normalised independently of the babble (unlike in previous chapters). In this work the level of the babble and the level of the SFX were separately normalised to -23dB^{LKFS} according to ITU-R BS.1770-2 specifications [323].

Additionally, to restrict the number of experimental factors, only overlapping, preceding non-speech objects were included. This was chosen as it represented a *'worst-case scenario'* – i.e. the more difficult condition as shown by Chapter 5. Additionally it allowed for direct comparison with the results of Chapter 6.

7.5.3 Multiple Signal to Background Ratio Paradigm

The stimuli were reordered into a multiple SBR paradigm to allow determination of the 50% intelligibility point on the psychometric curve for each participant. The range of s was based on the range of values selected by participants in Chapter 6: +14dB to -1dB.

To evenly span the range and give good granularity, 3dB steps were chosen. A visualisation of these can be seen in Figure 7.5. According to the Tillman-Olsen method for obtaining speech reception thresholds [346, 347], a minimum of one stimuli per 1dB is required. For example, in the QuickSIN, each step is 5dB and there are five target words. The QuickSIN also recommends that at least two lists (equating to 2 stimuli per step) be used to ensure accuracy [347]. For these reasons 6 stimuli per 3dB step was selected for this multiple speech to background version.

An additional 30 sentences were selected for use as a practice list, with only HP+SFX sentences. For the practise test a wider range of s, in 6dB steps spanning -29dB to -5dB was used. This wider range of s gave a trial criterion to determine whether the range of s in the main test would be too difficult. The practice list was not included in the phonetic analysis in Figure 7.3 or the smoothing in Section 7.5.3 as it was only used for practise and as a trial criterion.

Smoothing the pyschometric curve

When generating a multiple SBR test, the stimuli cannot be arbitrarily distributed. This is because, as noted by Wilson in previous work, when sentences for each SBR were selected arbitrarily from the original lists, both normal and hard of hearing listeners had irregular performances at adjacent s [190]. This is to say, participant's word recognition rate did not monotonically decrease with decreasing SBR as would be expected. This is due to the fact that when presented at the same , not all sentences are equally difficult to understand. The different sentence content and the fluctuating masker are the primary reasons for this.

To address this issue, Wilson generated recognition-performance data. This process involves determining which sentence/masker combinations have a higher average intelligibility and which had a lower average intelligibility with human listeners. Optimally, this would be achieved by gaining human evaluation scores for each of the stimuli at each of the possible s in the pre-determined range. However, as each stimuli/background ratio combination would require a different listener, Wilson deemed this impractical and instead generated the data using a reduced method. This method allocated each sentence to three possible s. The word recognition rate of these were then evaluated with a cohort of 48 older listeners with sensorineural hearing loss. Based on these results, the most difficult sentences on average would

be allocated to the lowest SBR of the set of three and the easiest to the highest s. Wilson only utilised the low predictability sentences as they represented the more challenging version of each keyword. The effect of this procedure is more homogeneous results at adjacent s and a smoother psychometric curve.

This still represents a very large undertaking and is also limited in its initial arbitrary allocation of sentences to groups of s. Wilson's choice to only utilise the low predictability sentences highlights that, for this task, the key variation we are addressing is differences in energetic masking. Given that it is only the energetic masking that needs to be evaluated, this task is perfectly suited for utilising a computational metric. By doing so, this would allow for all sentence/background ratio combinations to be evaluated in a resource efficient manner.

Methodology

Given the successful use of the GP in this work, this metric was selected as the most suitable [65]. As in Wilson's work, the recognition performance data was generated based on the LP sentences and applied to the matching HP sentence. It is true that the high and low predictability sentences would have slightly different GP due to the fluctuating masker, despite having the same keyword. However, it was deemed more important that the same keywords be assigned to the same SBR so that LP and HP measures are more reliably comparable. As the HP+SFX condition only had HP sentences, this procedure was repeated separately for these sentences.

For each SBR in the range -1 to +14dB, the GP for each keyword and babble pair was calculated. Each sentence had identical time alignment with the babble segment; only the level of the babble was altered. For each keyword, the GP was averaged across the the six SBRs ([-1, 2, 5, 8, 11, 14]). This average was then used to order the sentences from smallest average GP (the most difficult) to highest average GP (the easiest). In groups of six, these words were then allocated to each SBR from most difficult to easiest. The average GP per keyword, and their allocation into SBR increments, can be seen in Figure 7.6. In doing this, two pairs of words in the SFX + HP set were allocated incorrectly: *drain* and *frogs* were swapped, as were *breath* and *cheers*. As these were only swapped between adjacent SBRs, it was deemed to have minimal effect and they were maintained in these positions.

Once 6 LP (and its HP pair) and 6 HP+SFX sentences had been allocated to each SBR, they were pseudo-randomised into three lists which gave two of each sentence type at each level. These final forms of the lists, along with the sentences from the practise list, can be seen in Appendix D.

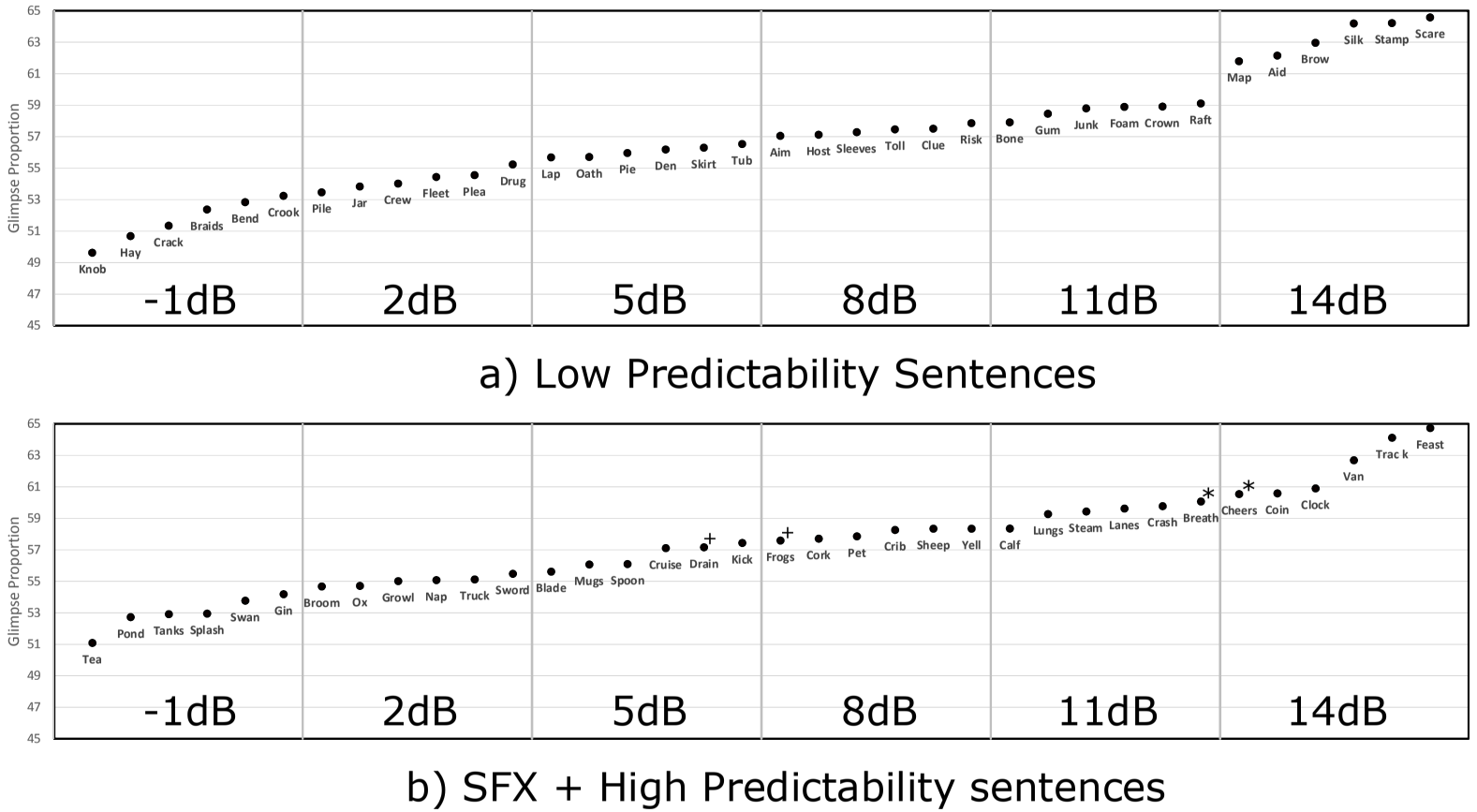


Figure 7.6 Mean glimpse proportion (GP) for each of the selected a) low predictability (LP) keywords and b) high predictability with redundant non-speech object (HP+SFX) keywords, ordered from smallest to largest and segregated into their speech to background ratio levels. The two incorrectly allocated pairs are noted by $^{+}$ and * above the corresponding keywords.

7.6 Chapter Summary

This chapter addressed the five key problems identified in the experimental methodology developed in Chapter 4 when implemented for hard of hearing cohorts in Chapter 6. These problems and their respective solutions were:

- *Limited characterisation of participants' suprathreshold hearing ability* – addressed by the inclusions of the Quick Speech Perception in Noise Test (QuickSIN), the 'Speech, Spatial and Qualities of Hearing Scale' (SSQ49) self-report survey and a measure of temporal fine structure.
- *Quality of the recording and speaker accent* – addressed by new high quality digital recordings with speakers using Received Pronunciation, termed the R^2SPIN.
- *Linked level of the redundant non-speech object to the multi-talker babble masker* – addressed by setting the redundant non-speech objects at a static level of -3dB relative to the speech throughout the experiment.
- *The single speech to background ratio yielding a word recognition rate*: addressed by reordering the new R^2SPIN into a multiple speech to background paradigm, which produces a speech reception threshold.

Analysis was also undertaken to ensure that the new multiple speech to background ratio format produced homogenous results at adjacent speech to background ratios and a smoother psychometric curve. This utilised a method adapted from the work of Wilson et al. work [190].

The developments of the methodology which are described in this chapter are utilised in Chapter 8 to complete this work's evaluation of the first research question.

Chapter 8

Linking hearing impairment and broadcast needs

8.1 Introduction

This chapter conducts this work's fourth and final study addressing the first research question:

What is the relationship between redundant non-speech audio objects and broadcast speech intelligibility, for normal and hard of hearing listeners?

This chapter implements the methodology developed in Chapter 4 and refined in Chapter 7. This methodology involves characterising participant's hearing utilising four measures: Pure Tone Audiometry [124], the Quick Speech in Noise (QuickSIN) test [331], a measure of Temporal Fine Structure (TFS) [336, 337], and the Speech, Spatial and Qualities of Hearing scale (SSQ49) [249]. The Re-recorded Revised Speech Perception in Noise test (R^2SPIN) is then utilised in its multiple speech to background form to determine how redundant, non-speech audio objects affect speech reception thresholds (SRT) of hard of hearing individuals. The relationship between this result and the different characteristics of hearing loss is then used to further elucidate the relationship between broadcast speech intelligibility and non-speech redundant sound objects.

The second part of this chapter takes the findings from all four studies in Part II, synthesising them into key conclusions for the first research question (stated above). In consequence of this, parameters to inform the subsequent work for the second research question are identified.

8.2 Study Four: Further evaluation of non-speech objects on intelligibility for hard of hearing cohorts

This section sequentially addresses each part of the experimental method for characterising hearing loss that was developed in Chapter 7. The descriptive statistics for each part of the experimental methodology are also given.

The tests described were all implemented with participants on a single visit to the University of Salford, with the exception of participants who completed the SSQ49 online. This visit lasted approximately six hours and included breaks between tests and a break for lunch to ensure participants were not fatigued. Participants were also given a tour of the Acoustic Laboratories and had the opportunity to engage with postgraduate researchers to make the experience more enjoyable and interesting.

The data collection was conducted by the author along with postgraduate and graduate staff who had received training from the author in the test they were conducting.[1] To allow the tests to be run in parallel, participants each undertook the tasks in a different order.

8.2.1 Participants

Recruitment

Participants were recruited through professional organisations (such as the Institute of Sound and Communications Engineers), community groups (such as University of the 3rd Age) and through University publicity. Participants were native English speakers and either identified as having mild to moderate hearing loss in their better hearing ear or were over the age of 50.

Demographics

There were 18 participants who took part – 14 males and 4 females. Participants ranged in age from 53 to 95, with a median age of 65. Only two participants self-identified as being musicians (amateur). All spoke English as their first language and all bar one were resident in the UK.

[1] The staff who performed data collection for this study were: Dr. William Bailey, Dr. Lara Harris, Dr. Ben Shirley, Zuza Podwinska, Philippa Demonte and Georgia Shirley. Staff were either employed by the University of Salford or paid for their time.

Table 8.1 Counts for each hearing loss severity based on low and high frequency pure tone averages, with the symmetry of loss noted.

Hearing Loss Severity	PTA (counts)	HFPTA (counts)
Normal	1	1
Asymmetric Mild	3	1
Asymmetric Moderate	–	1
Asymmetric Severe	1	1
Symmetric Mild	5	3
Symmetric Moderate	7	6
Symmetric Severe	1	3
Symmetric Profound	–	2

8.2.2 Hearing loss characterisation

Pure tone audiometry

Methodology Audiograms were performed according to the BSA Recommended Procedure [124] utilising either a Kamplex r27a or Kamplex KLD21 Diagnostic Audiometer. Audiometric thresholds for the frequencies 0.25Hz, 0.5Hz, 1kHz, 2kHz, 4kHz, and 8kHz were obtained. Two pure tone averages (PTA) were calculated; PTA averaged over the frequencies 0.25Hz, 0.5Hz, 1kHz, 2kHz and 4kHz and a high frequency version (HFPTA) averaged over 2kHz, 4kHz and 8kHz. Thresholds were tested up to 110dBHL and if a subject had no response at this level, that frequency was reported as a 'no response' (*NR*). When calculating average values, the BSA recommended *NR* value was used: – 130dB .

Descriptive statistics Counts for the number of participants with each degree of hearing loss can be seen in Table 8.1. These were calculated according to the methods described above and the definitions of hearing loss severity outlined by the British Society of Audiology [124]. Whether this loss was symmetric or asymmetric is also noted. If hearing was normal in one ear and impaired in the other, this was defined as asymmetric hearing loss. If the hearing in both ears was impaired, but not equally, this was defined as symmetric hearing loss and the severity was defined by the hearing loss in the worse ear.

As would be expected for older listeners, their PTAs were skewed towards more severe loss at high frequencies. The majority of participants had symmetric loss. All participants exhibited some degree of hearing loss as per these definitions. One participant had normal low frequency PTAs, though had a mild high frequency loss, and another participant had normal high frequency PTAs with mild low frequency loss.

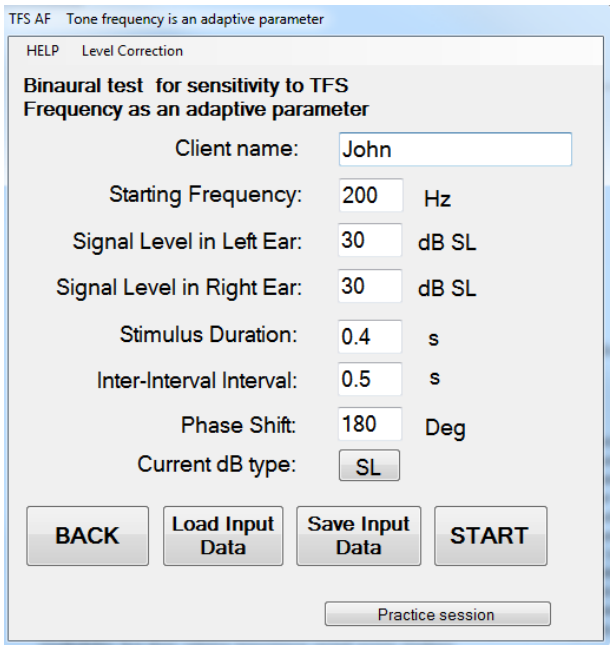


Figure 8.1 Settings utilised for the Temporal Fine Structure - Adaptive Frequency test [337, 336]. The starting frequency was 200Hz and the signal was set to 30-40dB more intense than participants' hearing thresholds, in each ear, as measured in Section 8.2.2.

Adaptive Frequency Test of Temporal Fine Structure

Methodology TFS was measured utilising the TFS-AF test developed by Füllgrabe and colleagues [336, 337]. The settings utilised can be seen in Figure 8.1 and the level was set at 30-40dB above participants' hearing loss. The test was reproduced over Sennheiser HD 800 open backed headphones in an acoustically insulated room. Level adjustment was applied for each participant based on the results of their pure tone audiometric thresholds at 0.25kHz, 0.5kHz, 1kHz, and 2kHz to ensure that all frequencies were perceived at the same loudness by all participants. The starting frequency was set to 200Hz. An example output is given in Figure 8.2.

Descriptive statistics One participant was unable to complete the task as they were unable to perceive the difference at the initial setting due to the severity of their hearing loss. Their results were omitted.

The frequency at which participants could no longer discriminate interaural phase differences ranged from 41.8Hz (SD = 0.17) to 1687.9Hz (SD = 0.11) with a median value of 1042Hz. This highlights the wide spread of values which were elicited from participants.

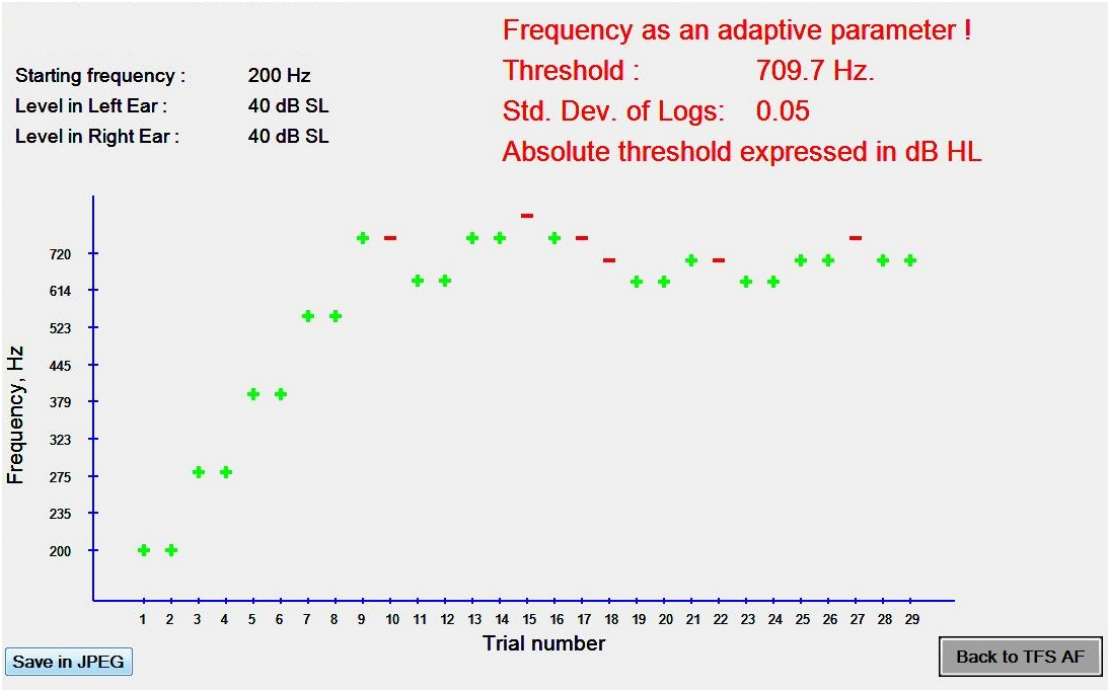


Figure 8.2 Example output of the Temporal Fine Structure - Adaptive Frequency test [337, 336], with green crosses noting correct responses at each frequency and red minuses denoting incorrect responses. The final calculated threshold and the standard deviation is given in the top right in red text.

QuickSIN

Methodology The QuickSIN test was installed on the Kamplex r27a Diagnostic Audiometer and presented to both ears at the same level over audiometer's headphones. One practice list (Practice List A; Track 21) was presented first to all participants. The evaluated portion of the test consisted of two of the main lists (Track 2; List 4 and Track 1; List 3), with half the participants receiving Track 4 first and half receiving Track 3. It is recommended that for improved accuracy two or more lists should be averaged [347]. For this reason, participants completed two lists and the reported results are an average of the results of these lists.

A nominal presentation level of 65dBSPL was used and, during the practice list, participants could request the level be increased by up to 10dBSPL. Participants were asked to write down the whole sentence and the progression of the sentences was paused by the researcher to allow each participant adequate time to write. Participants were also given the option of having the researcher scribe for them if they preferred.

SNR loss was calculated using the formula:

$$\text{SNR Loss} = 25.5\text{dB} - (\text{Total Correct}) \tag{8.1}$$

Table 8.2 Severity levels of speech to noise ratio (SNR) loss and number of participants at each level, calculated using Eqn. 8.1.

SNR Loss	Count	SNR Loss Severity
< 0dB	1	Normal
0-3dB	6	Near Normal
3-7dB	2	Mild
7-15dB	7	Moderate
>15dB	2	Severe

the derivation of which can be found in the QuickSIN Manual [347] and is based on the Tillman-Olsen method for obtaining spondee thresholds [346].

Descriptive statistics SNR loss ranged from -1.5dB SNR to 22.5dB SNR, with a median value of 6.5dB SNR. The QuickSIN manual provides an interpretation of the resulting SNR loss in terms of severity. The number of individuals at each severity level is given in Table 8.2.

We can see that whilst only one participant had what the QuickSIN deems 'normal' speech in noise performance, one third of participants had near normal performance. The majority of participants had moderate SNR loss.

Speech, Spatial and Qualities of Speech Survey and self-reported Hearing Loss

An in-depth description of the SSQ can be found in Chapter 3. The questions in the full SSQ49 version, which differs from the SSQ12 version utilised in Chapter 3, can be found in Appendix E. In addition to the SSQ49, participants were asked whether they had hearing aids fitted and what their hearing aid usage patterns were.

Methodology Participants had the choice to complete the SSQ49 either online or in person when they arrived at the University. Online implementation used the same platform as Chapter 3: [onlinesurveys.co.uk]. Included in the online delivery were the survey questions from 8.2.3 on broadcast media experiences.

Descriptive statistics Nine participants had hearing aids fitted in at least one ear (5 bilateral, 4 unilateral) and of those, five identified as using these devices regularly. These participants indicated that they had had their assistive hearing devices fitted for periods ranging from 2 to 16 years and used a variety of aids. Seven participants reported suffering from Tinnitus.

The self-identified severity of hearing loss is summarised in Table 8.3

Table 8.3 Individuals' self-reported hearing loss severity, with number of participants identifying with each level of hearing loss given.

Hearing Loss Severity	Counts
Normal	3
Mild	5
Moderate	9
Not sure of severity	1

Any omitted values from the SSQ49 were imputed to the mean of that individual's scores for that section of the survey (Speech, Spatial or Qualities). One question was accidentally omitted from the Qualities of Hearing section (question 14), meaning that in this implementation the 'Qualities of hearing' section only contained 17 questions.

It can be clearly seen from Figure 8.3 that individuals' perceptions of their spatial hearing ability and the quality of their hearing was high than their perception of their ability to hear speech in noise. For this reason, we can see that the overall SSQ49 value is skewed towards ratings of higher ability.

8.2.3 Broadcast speech intelligibility evaluation

This section outlines the two measures used to evaluate an individuals' ability to understand broadcast speech. The first is a qualitative assessment, using the ten self-report questions, on ease of understanding speech on television and subtitling usage (termed TV10) developed in Chapter 3. The second is the quantitative assessment method developed in Chapter 7 using the multiple speech to background ratio version of the R^2SPIN. The implementation and descriptive statistics for these measures are given, along with a preliminary discussion of the results.

Survey

The questions developed in Chapter 3 to characterise individuals hearing (dis)ability in various television based listening scenarios were also used here. These questions can be seen in Appendix B and in Table 8.4.

Methodology The TV10 questions were included in a survey along with the questions from the SSQ49. Participants could choose either to complete online in their own time or complete the survey on the day they visited the University.

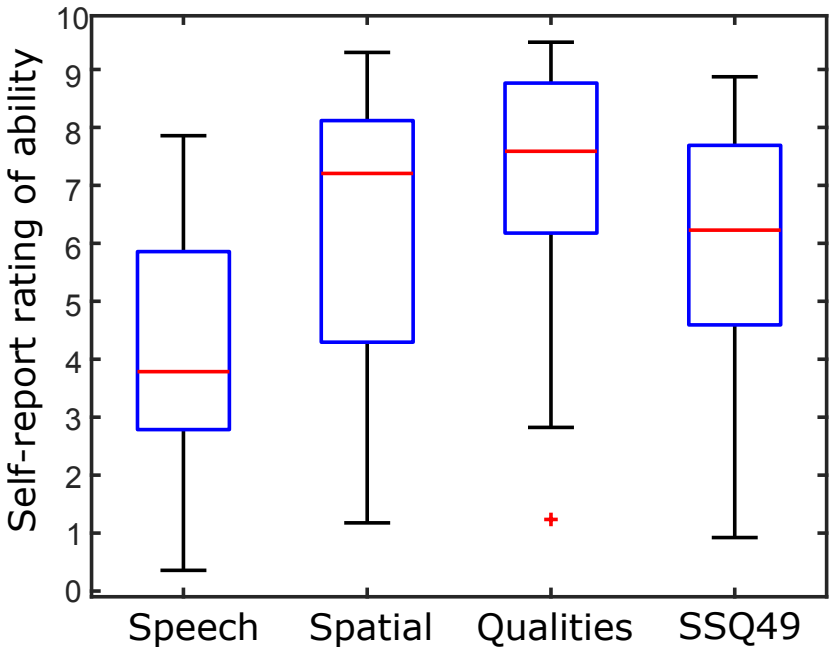


Figure 8.3 Box plot of each individual's mean rating for each section of the Speech, Spatial and Qualities of Hearing Scale respectively, as well as their overall mean rating across all the sections (denoted as SSQ49). *Median values are given in red, with the blue indicating the interquartile range (between the 25th and 75th percentile), the whiskers indicate the minimum and maximum values and outliers denoted by a red cross.*

Descriptive statistics The types of content most commonly watched by participants were 'News and Current Affairs' (14 participants) and 'Documentary' (13 participants), followed by 'Drama and Soaps' (12 participants). The number of hours per day spent watching television can be seen in Table 8.4. It can be seen that the majority watch 1–4 hrs but few watch more or less than this.

When asked which genre participants found the easiest to understand the speech in, 15 participants identified 'News and Current Affairs', followed by Documentary (13 participants). Only two participants identified 'Drama and Soaps' and 'Films' and only one identified 'Lifestyle, Music and Food' and 'Comedy'. When asked which sounds help to follow a plot in a drama, almost all participants identified dialogue (17 participants) and five participants identified foreground sounds; two participants identified background sounds and one identified music.

Participants gave a wide range of responses to question TV1, spanning almost the whole range of values and the median values was 6. Questions TV2, TV3, TV5, TV6, TV7 and TV8 are also considered scenario-based questions (see Table 8.5. From this we can see that the most difficult scenario was TV3, which relates to a panel show with background laughter. The easiest, as in Chapter 3 was TV5, a news reporter in a quiet studio. Most participants

Table 8.4 Self-reported hours of television watched per day, and number of participants for each response.

Hours	Number of Participants
Less than 1 hr	1
1-2hrs	7
3-4hrs	8
More than 4hrs	2

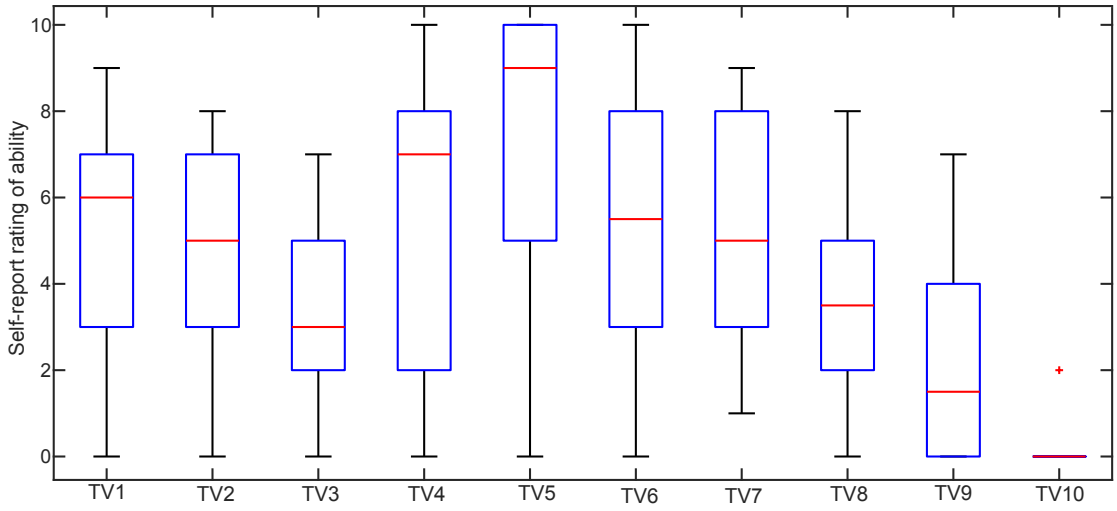


Figure 8.4 Box plot of responses to the TV10 questions, showing the median value for each question (red), the interquartile range (blue) and the whiskers denoting the range of values. Corresponding questions can be seen in Table 8.5.

indicated that hearing what is said in dramatic content requires a great deal of effort, with a median value of 3.5.

Four participants indicated that they never used subtitles (TV4) whilst those who did use them, tended to use them regularly as seen by the median value of 7. Few participants watched content in another language, with a median value of 1.5 (TV9). Only one participant reported watching signed content but rarely (TV10, noted as an outlier in Figure 8.4).

R^{2}SPIN

Methodology The background and speech stimuli were co-located and presented from a Genelec 8030A Studio Monitor mounted at a height of 1.1 m and the listener was situated 1.1 m from the speaker. Half of the participants completed the task in a listening room meeting the ITU-R BS.1116-1 standard for listening tests [324] whilst the other half completed it in an acoustically treated but not isolated room. The stimuli were presented at a sound pressure

ID	Survey Items	Scale (0 → 10)	Scenario-based
TV1	Generally, how difficult do you find it to understand speech on television?	Very Difficult → Very easy	Yes
TV2	A character is speaking but they are not on screen. How easily can you understand the speech without seeing the character's face?	Not at all → Perfectly	Yes
TV3	You are watching a panel show and one of the panellists is speaking whilst the studio audience laughs and cheers. How easily are you able to understand the panellist's speech?	Not at all → Perfectly	Yes
TV4	How often do you use subtitles?	Never → Always	–
TV5	A news presenter is reporting from a quiet studio. Without using subtitles, how easily can you understand the speech?	Not at all → Perfectly	Yes
TV6	You are watching a scene on television which has the sound of clinking glasses, music and people talking in the background. Can you make out the different sounds?	Not at all → Perfectly	Yes
TV7	You are watching a nature documentary. The narrator is speaking with the constant sound of a waterfall in the background. Can you follow what the narrator is saying?	Not at all → Perfectly	Yes
TV8	How much effort do you require to hear what is being said in a television drama?	A lot of effort → No effort	Yes
TV9	How often do you watch television programs which are not in your native language?	Never → Always	–
TV10	When sign-interpretation is available, how often do you watch sign language interpreted programming?	Never → Always	–

Table 8.5 Quantitative 11 point Likert scale questions on television experience, showing question ID and range of values, developed in Chapter 3 and corresponding to the boxplot shown in Figure 8.4. The scenario-based questions which are used in analysis in Section 8.2.5 are noted.

level of 69 dB(A), measured at the listening position as previously used in Chapters 5 and 6. The three main lists were presented to the participants in a pseudo-random order. The practice list was presented first, to allow participants to familiarise themselves with the task.

Calculating the speech reception threshold The two primary ways in which the speech reception thresholds (SRT) can be calculated for multiple speech to background ratio tests (with monotonically decreasing rather than adaptive stimuli) are averaging over the speech to background ratios (as used in QuickSIN [347]) or by fitting a psychometric function. Whilst fitting a psychometric function to the resulting data provides the best test-retest reliability [348], this approach relies on a number of assumptions which are often not met [349]. The first assumption is that the response data is monotonically decreasing. Whilst every effort was made in Section 7.5.3 to create stimuli which would elicit monotonic responses, this cannot account for individual variations which affect keyword recognition, such as lapses in attention. The second is the assumption of binomial distribution with independent trials. The use of single keywords per stimuli does increase the independence of the trials. However, the effect which results from learning the pattern of the stimuli and acclimatising to the babble noise and speaker's voice mean that the assumption of independent trials is still violated.

For this reason, an approach based on the mean speech to background ratio was adopted. This was developed using the same approach as the QuickSIN [347]. This consistency between SRT measures also allows more reliable comparison between individual's scores on both experimental tests.

To calculate the SRT, the following equation was derived:

$$\text{SRT} = 15.5\text{dB} - \left(\frac{\text{Total Correct}}{2}\right)$$

This takes the highest step value (in this case 14dB) and adds the value of half a step, 1.5dB, giving a value of 15.5dB. The total number of correct keywords are then averaged as there are 6 keywords for each 3dB step (unlike in the QuickSIN, which had only the minimum value of one keyword per dB step). Furthermore, the QuickSIN value was offset from the average speech to noise ratio collected for normal hearing listeners completing the QuickSIN test – 2dB. Given that no normative values for the multiple speech to background version of R^2SPIN with SFX exist, such an offset was not used.

Descriptive statistics The data was investigated to determine whether any participant had demonstrably found the task too easy (results saturated at the floor of the psychometric curve) or too difficult (saturated at the ceiling). Evaluation of the latter was facilitated by the practice list with an extended range of speech to background ratios; those with SRTs from the practice test which were well beyond the range of the main test could be reliably excluded. The SRTs in the practice list ranged from 2dB (lowest possible value) to 29dB. The two participants who scored a SRT of +20dB on this list were omitted. The participant who scored 19dB on the practice test, but had SRTs well within the main test's range was

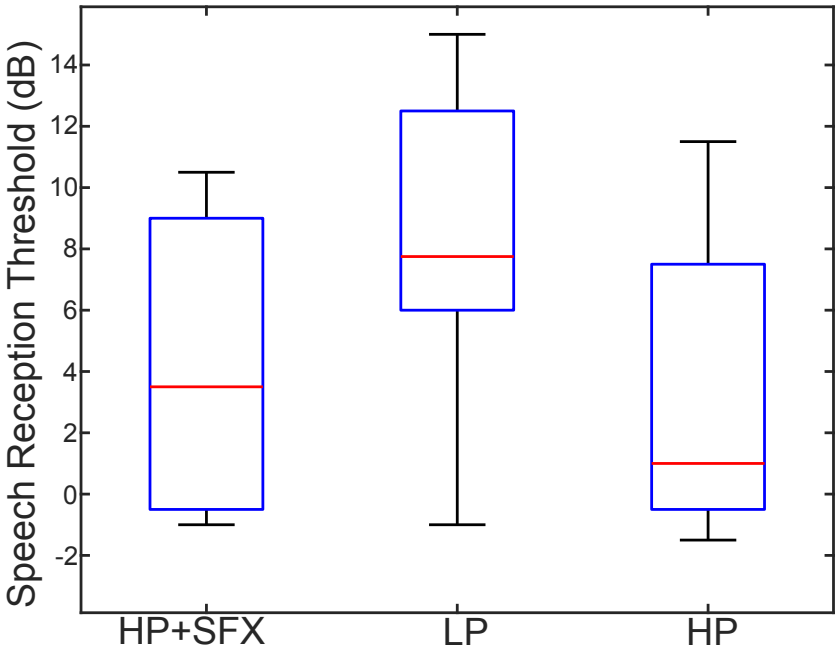


Figure 8.5 Box plot of the speech reception threshold in each of the three conditions in the R^{2}SPIN, showing the median value for each question (red), the interquartile range (blue) and the whiskers denoting the range of values.

retained. Those who found the test too easy were also excluded. This was defined as those whose SRT was -1dB or less (i.e. those who got three or fewer keywords wrong). A further six participants were excluded for this reason. This left a cohort of 10 participants for the following analysis.

A box plot of the participants who successfully completed the task can be seen in Figure 8.5. For the LP condition, SRTs ranged from -1 dB to 15dB, with a median value of 7.75dB. For the HP+SFX condition, participants demonstrated SRTs in the range -1dB to 10.5 dB with a median value of 3.5dB. For the HP condition, the SRTs had a slightly broader range from -1 dB up to 11.5dB, again with a median value of 1dB. From the median values, both the LP and HP conditions are skewed towards lower values. The median of HP+SFX was roughly central, demonstrating a much less skewed spread. Interestingly, as seen in Chapter 6, some participants benefited from the presence of additional non-speech content, whilst others found it degraded word recognition.

These results were further investigated using the SRT as the response variable in a repeated measures ANOVA (implemented using ezANOVA from the EZ package in R [350]). The assumption of sphericity was assessed using Mauchly's test. The data was found not to violate this assumption ($[p > 0.05]$) and subsequently, no adjustments were applied. The main effect was found to be significant ($[F = 23.2, p < 0.001]$). Post-hoc examination

utilising contrasts to explore the pairwise comparisons showed that the conditions LP & HP and conditions LP & HP+SFX were significant at the level $[p < 0.001]$ and $[p < 0.01]$ respectively. There was no significant difference between the HP and HP+SFX conditions. These contrasts were conducted using a pairwise t-test with a Bonferroni correction.

8.2.4 Preliminary discussion

The significant difference between keyword recognition for LP and HP sentences is replicated here as in other results in this work (Chapter 5 and 6) and the established literature [319, 190]. This reaffirms Chapter 7's results which showed that the new version of the R^2SPIN stimuli is still valid for evaluating the effect of predictability on speech recognition in noise.

A large spread of SRT values was seen across all conditions even when those for whom the task was too difficult or too easy were excluded (see Figure 8.5). Given the variance in SFX benefit for high predictability speech in Chapter 6, it was not unexpected that no significant difference would be seen in the results here. However, the high number of excluded participants indicate that a larger range of speech to background ratios for the main test would allow for more useful data to be collected in future. Furthermore, an extended range practise list which also identifies those for whom the task is too easy would improve reliability should participants still need to be excluded.

8.2.5 Analysis of results

This section aims to determine the answer to the first research question for high predictability speech perception in hard of hearing cohorts. This is achieved in a similar manner to that of Chapter 6 where improvement (or degradation) due to the presence of SFX is calculated for each individual. Correlation analysis is then used to determine which aspects of hearing loss relate to this improvement most strongly and may help establish a hypothesis for the causal mechanism behind the utility of SFX for hard of hearing listeners. The relationship between the two measures of broadcast intelligibility, the TV10 and the HP+SFX SRT, are then explored. This aims to determine whether self reported perception of broadcast speech understanding relates to the controlled HP+SFX SRT. If so, the TV10 may be used to predict the required speech to SFX to background ratios needed for individuals, based only on self-reported data.

Correlation analysis

To explore the relationship between the measured aspects of hearing ability and the usefulness of redundant non-speech audio objects (SFX), a correlation analysis was performed. As in

Table 8.6 Partial correlation using Pearson's correlation coefficient, between each measure of hearing loss and improvement in intelligibility with redundant non-speech objects (SFX), controlling for all other measures of hearing loss.

	Hearing aid	TFS	PTA (Better ear)	SNR loss	$SSQ49_{mean}$
Improvement	-0.34	0.73	0.84*	-0.73	0.75

$*p < .05$, $**p < .01$, $***p < .001$

Chapter 6, a measure of percentage 'improvement' is used and this is replicated below for convenience.

$$\text{Improvement} = \frac{\text{WRR in HP+SFX} - \text{WRR in HP}}{\text{WRR in HP}}\% \tag{8.2}$$

Word recognition rate (WRR) is used here rather than SRT, as it does not require the offset described in Equation 8.2.3. The relationship between this improvement variable was explored with the following hearing loss measures: TFS, PTA in the better hearing ear, SNR loss (measured by QuickSIN), the average SSQ49 score, and a binary value indicating whether the participants used a hearing aid. The number of measures explored had to be restricted to ensure that the reduced participant pool did not become overdefined. It is for this reason that the single SSQ49 average, rather than the average of the individual sections of the SSQ49, was selected. The participant for whom a TFS measure could not be obtained had their TFS imputed to the mean of the cohort.

All variables were shown to be normally distributed using the Anderson-Darling Test, with the exception of the hearing aid usage variable which is a two level categorical variable. As one two-level categorical variable is permissible in the calculation of Pearson's correlation coefficient [270], this test was used. A two-tailed partial correlation was performed as it is highly likely that the different measures of hearing loss are themselves correlated. The use of a partial correlation allows for these effects to be controlled for to reveal any factors which independently predict improvement. This was performed using partialcorr() in the Matlab Statistics and Machine Learning Toolbox.

Results The results of the correlation analysis can be seen in Table 8.6. This shows that an individual's PTA in their better hearing ear is a significant predictor of the usefulness of SFX in high predictability speech (when other hearing loss characteristics are controlled for). This has a Pearson's R which indicates a strong correlation. This mirrors the result found for low predictability speech in Chapter 6; the correlation coefficient which is produced is also of a very similar magnitude to the relationship found in Chapter 6. This indicates that PTA in

an individual's better hearing ear is a strong predictor for SFX utility in both low and high predictability speech.

Relationship between TV10 and SFX SRT

The results in Section 8.2.5 and 8.2.3 indicate that there is measurable difference in speech recognition performance when SFX are present and when they are absent. This section now explores how this measurable effect relates to an individual's self-reported perceptions of their broadcast speech understanding.

In this analysis, only seven of the ten TV10 questions are used: those relating to behaviour (e.g. *TV9 - How often do you watch television programs which are not in your native language?*) are omitted. The reason for this is that both TV9 and TV10 are significantly different in their responses from the other questions (see Figure 8.4: medians lying outside the interquartile range of the other questions). For TV4, which is not significantly different to all other questions, exclusion was based the results it was shown in Chapter 3 that subtitle usage behaviour is influenced by many factors of which television speech understanding is only one. This leaves seven remaining questions which were averaged for each individual.

The SRT value for condition HP+SFX is used here, rather than the improvement measure used in Section 8.2.5, to link the TV10 results to practical dB values. A scatterplot and correlation analysis were performed to investigate this relationship. As the variables are normally distributed, Pearson's correlation coefficient is again used.

Results

A scatterplot of each individual's average response to the seven scenario based questions from TV 10 and their SRT for the HP+SFX condition are shown in Figure 8.6, with $[r = -0.79, p > 0.01]$. From the Pearson r value, we can see a strong negative correlation. This indicates that the higher an individual rates their ability to hear speech in TV scenarios, the lower the speech level they require to understand 50% of what is said (in the presence of noise).

To determine whether the results from the TV10 questionnaire are simply acting as a proxy for an individual's perceived speech perception ability in all scenarios, a partial correlation analysis was performed. The results of this can be seen in Table 8.7. It can be seen that when general self-reported speech understanding is controlled for, the relationship between self-reported television speech understanding and an individuals' HP+SFX SRT is still significant and, in fact, increases. This echoes the results seen in Chapter 3 that showed independence between general speech understanding and speech understanding on television.

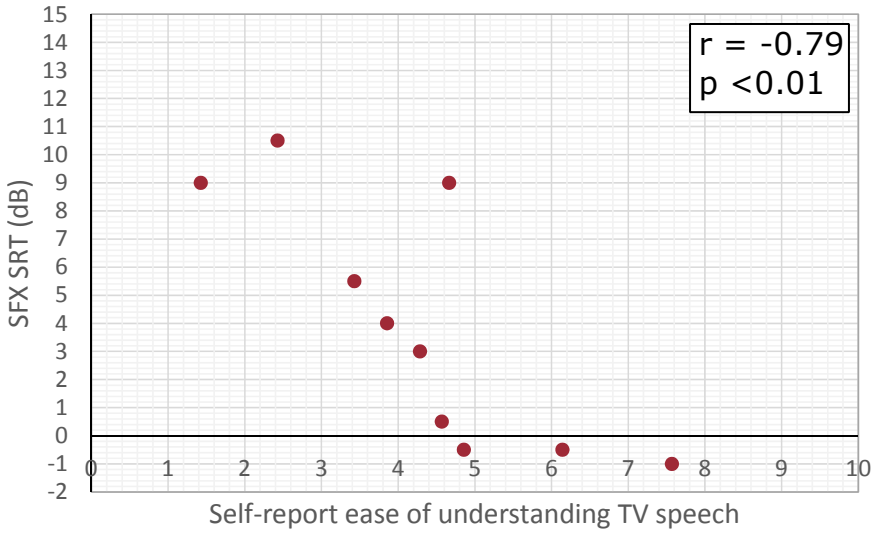


Figure 8.6 Scatterplot of individual's average response to the scenario specifics questions from the television experience questions (TV10), in Table 8.5 and individuals' speech reception threshold (SRT) for the high predictability speech with redundant non-speech object condition (HP+SFX).

Table 8.7 Partial Pearson correlation between individual's average response to the scenario specifics questions from the television experience questions (TV10), in Table 8.5, average response on the Speech, Spartial and Qualitative of Hearing Scale (SSQ49) and their speech reception threshold (SRT) for the high predictability speech with redundant non-speech object condition (HP+SFX).

	TV10 (scenario questions)	HP+SFX SRT
$SSQ49_{mean}$	0.62	0.43
TV10 (scenario questions)	–	-0.82**

$^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$

8.3 Answers to the first research question

Throughout Part II of this work, there have been numerous findings which address the first research question:

What is the relationship between redundant non-speech audio objects and broadcast speech intelligibility, for normal and hard of hearing listeners?

This section discusses these findings and the conclusions, which Part III addresses the second research question, will build upon. Additionally, the open source dataset this work has produced, and how they may be used to undertaken future work, is described.

8.3.1 The relationship for normal hearing

Studies One and Two, described in Chapter 5, concluded that the presence of redundant non-speech audio objects improved the intelligibility of speech in noise. Based on the controlled factorial study design and quantification of key confounding factors (like the effect of critical listening experience), this conclusions is drawn with a reasonable degree of certainty.

Study One 5.2 demonstrated that the presence of SFX improved speech recognition even when the SFX overlapped some of the speech. Greater improvement could be achieved when the SFX followed the keyword, as shown by Study Two. In this case, the improvement due to the semantic context of the sentence (predictability) and the improvement due to the non-speech context became additive. Through modelling, it was determined that the effect of overlapping the SFX with some of the stimuli's speech was equivalent to reducing the level of the speech by 4.5dB. This demonstrates that use of non-speech audio context in an overlapping scenario is more difficult than the scenario where the SFX does not overlap the speech.

Through the use of a computational intelligibility metric, the glimpse proportion, it was shown that the increased energetic masking due to the SFX contributed to this difference. Furthermore, it was hypothesised that the overlapping scenario requires more cognitive effort. This effort would be spent on attention switching between the speech, SFX and babble in order to successfully segregate audio elements and parse their meaning [328]. However, conclusions on this effect were not able to be drawn from the collected data.

8.3.2 The relationship for hard of hearing cohorts

Study Three and Four described in Chapters 6 and Section 8.2 respectively indicate that the relationship between redundant non-speech audio objects and broadcast speech intelligibility are different for various hard of hearing listeners. For this reason, in both these studies, a measure of 'improvement' has been used to indicate whether the presence of SFX aids a listener's ability to understand speech in noise, makes no difference, or actively degrades this ability. For both high and low predictability speech, this improvement value represents percentage increase in keyword recognition compared with high or low predictability speech without SFX. Thus, it is also referred to as the utility of the SFX for a particular individual.

Given these observed difference between listeners, and the consistency observed between the normal hearing cohort, it was hypothesised that a particular aspect of hearing loss caused the differences in this relationship for hard of hearing listeners. However, as characterisation of all aspects of hearing loss is not possible, a correlational rather than casual study of the effect was undertaken. The aim of this was to determine which factors of hearing loss most

strongly relate to the utility of SFX for an individual listener. This would provide novel understanding about the relationship. Additionally, by narrowing down the aspects of hearing loss related to this effect, it serves as the foundation for future mechanistic studies.

The results of Study Three and Four demonstrate that there is a strong and significant relationship between PTA hearing loss in an individual's better hearing ear and the utility of SFX on speech understanding in noise. This means that as the average value of the lowest audible sounds at 0.25Hz, 0.5Hz, 1kHz, 2kHz and 4kHz in an individual's better hearing ear increases, the measured utility of SFX for them decreases. The reliance on an individual's better hearing ear is hypothesised to be due to the co-location of the stimuli, preventing spatial release from masking.

Study Three conclusively demonstrated this effect only for low predictability speech. It also demonstrated that the level of utility approaches that exhibited for normal hearing listeners as audiometric thresholds approached zero. This suggests that the utility of SFX may be a continuous variable from high positive levels for normal hearing listeners through to low, negative values for more severely hearing impaired individuals. Chapter 6 also hypothesised the existence of this relationship for high predictability speech but was unable to give strong evidence for it. Limitations in Study Three's method meant that, for some participants, the SFX may have been indistinguishable from the noise or simply inaudible. Furthermore, it is likely that this relationship was confounded by the additional processing of semantic context in the high predictability stimuli and the relatively limited information about the participant's hearing loss.

The design of Study Four in Chapter 7 addressed many of these issues including setting the SFX at a static level independent of the background noise and additional tests to characterise the participants' hearing loss. This allowed Study Four to demonstrate that this strong, significant relationship also exists for high predictability speech. This was achievable as the greater characterisation of hearing loss allowed for those factors to be controlled for in the correlational analysis.

The controlled level of the SFX in Study Four indicates that for those participants who had degraded speech recognition in the presence of SFX this was not due to, or at least not solely due, to the audibility of the SFX. This gives weight to the hypothesis put forth in Chapter 6 that attention and stream segregation ability may be playing a significant role in determining the usefulness of redundant non-speech audio objects. Characterisation of an individual's stream segregation ability for co-located stimuli in future work would allow for this hypothesis to be tested.

From Study Three and Four we can conclude that an individual's PTA in their better hearing ear predicts the utility of SFX in speech understanding, for both high and low

predictability speech. However, it is still unknown whether their higher audiometric threshold [351] is the cause of this reduced utility or whether it is a proxy measure for the real cause.

8.3.3 Linking hearing impairment and broadcast needs

The results of these four studies provide additional pragmatic results that can be utilised to develop accessible audio strategies, even in the absence of fully characterising the cause of varied SFX utility in hard of hearing cohorts. This information can be used to underpin the work in Part III, which addresses the second research question:

Can a system be designed to allow end-users to control the balance between audio objects for dramatic content which is simple to use and preserves comprehension?

The first practical result is that the effect of redundant non-speech objects on broadcast speech intelligibility is not uniform for hard of hearing individuals. Not only is the utility of these audio objects different in magnitude for different listeners, for some listeners it actively degrades intelligibility. This reaffirms the conclusions of Chapter 2 which identified that for an accessibility strategy to be effective, it must be personalisable.

The second practical output is the SRTs obtained using the R^2SPIN in a multiple speech to background ratio paradigm. The SRT is based on identifying 50% of keywords, which is clearly insufficient for understanding television content. However, this value is for a scenario where the stimuli are co-located and there is no visual context. Furthermore, this represents the scenario at which the SFX is temporally overlapping the speech which is not the usual case for television audio production. As such, we can treat these values as the worst-case scenario and, in normal scenarios, will likely provide much more than 50% intelligibility.

From these worst-case scenario values, it is suggested that the majority of those with mild to moderate hearing loss require a speech level which is in the range -1dB and +14dB, relative to the background noise, in the presence of a -3dB SFX. Any accessible audio strategy must then be able to adjust the background to a level up to 15 dB less than the speech (or lower for those with higher levels of hearing loss). At that level, a redundant non-speech object at -3dB is then acceptable, at least for those with mild to moderate hearing loss.

Finally, Section 8.2.5 showed that the average results of an individual's self-reported understanding of TV speech predicts their SRT for speech in background noise with SFX. This indicates the individuals' perceived speech perception ability for television speech is representative of their needs. This further motivates a personalised approach to accessible audio, indicating that individuals understand their own media access needs.

8.3.4 University of Salford Accessibility and hearing Impairment Database (USAID)

A limited number of databases exist that document the relationship between different clinical measures of hearing loss and speech intelligibility in everyday scenarios [352, 353]. Only one open-access data set contains survey data which links self-reported hearing loss with experience of television speech [15]. No open-access databases exist linking clinical measures of hearing loss with individuals' experience of broadcast media. To address this deficit, and to ensure that the rich data-set collected in this chapter has value beyond this current study, the anonymised data-set has been made available to other researchers under a creative commons license.

8.4 Part II Summary

This chapter describes the final of four perceptual studies designed to understand and define the relationship between redundant non-speech audio objects and broadcast speech intelligibility. This chapter began by describing the implementation and descriptive statistics for each part of the experimental methodology developed in Chapter 7. Correlation analysis between the different characterisations of hearing loss and the improvement in speech recognition in the presence of SFX was then conducted. The findings of this study, along with the findings from Chapters 4 – 7, were then discussed. From this, the following conclusions about the relationship between redundant non-speech objects and broadcast intelligibility were drawn:

- For normal hearing listeners, redundant non-speech audio objects aid intelligibility and:
  - are strictly additive with the effects of semantic context when the SFX does not overlap the preceding speech.
  - still aid keyword recognition when they overlap preceding speech, but degrade signal-level intelligibility the equivalent of reducing the speech level by 4.5dB.
- For hard of hearing listeners, redundant non-speech audio objects aid some listeners but degrade intelligibility for others:
  - The degree of utility a redundant non-speech object provides is predicted by an individual's pure tone average in their better hearing ear.

Additionally, this work has demonstrated the benefit of a mixed methods approach when evaluating intelligibility which allows the balance of repeatibility, representativeness of hearing loss, and ecological validity. It has also published two resources, the re-recorded

Revised Speech in Noise test (R^2SPIN) stimuli and the University of Salford Accessibility and hearing Impairment Database (USAID), to facilitate further research in this area. Finally, Part II has derived additional pragmatic results which will underpin the work in Part III. These results reaffirm the need for user-driven personalised accessibility strategies which provide a minimum range of 15dB between speech and background sound in the presence of -3dB SFX.

Part III

The Engineering

Chapter 9

Narrative Importance: An ecological approach to accessible audio

There seems to be a relative unanimity that narrative does not merely list what happens, but that it brings out or creates meaningful connections between events or experiences, thereby rendering them (at least partly) intelligible.

– Meretoja in "Narrative and human existence"[354]

9.1 Introduction to Part III

Part III of this work conducts four qualitative user-experience studies to address the second research question developed in Chapter 3:

Can a system be designed to allow end-users to control the balance between audio objects for dramatic content, which is simple to use and preserves comprehension?

This chapter first addresses the challenge of developing a categorisation for audio objects in television sound which is both intuitive to production staff and easily leveraged by the end-user for personalisation. An approach, inspired by work in the field of audio description and based on Gibson's theory of affordances in visual perception, is then developed. This new approach, termed 'Narrative Importance', is outlined and an object-based audio personalisation system based on this approach is then proposed. A prototype of a proposed system is then employed for Study One, the first of the user-experience evaluations. The result of this evaluation indicates that the prototype is considered by a cohort of older and hard of hearing individuals as a positive direction for accessible broadcast audio. A number of areas where the prototype can be improved are also identified.

Based on this prototype, Chapter 10 presents two further studies evaluating whether the proposed narrative importance [1] approach is commensurate with production staff's prioritisation of audio objects, and could be feasibly integrated into their workflows. The first of these studies conducts an ethnographic case study. Study Three builds on this with a larger cohort, exploring whether the results of the case study translate to a broader population of production staff. Finally, Study Four in Chapter 11 brings together the end-user and production evaluations in a large scale, public trial of the narrative importance approach, with an ecologically valid prototype system and broadcast content.

9.2 Prioritising sounds

9.2.1 Finding a common language

Part I of this work highlighted that the current provision of access services for those who are hard of hearing are not sufficient. Chapter 3 showed that dramatic content in particular presents a challenge, due to the complex soundscapes it presents. Whilst a large proportion of audiences face challenges engaging with broadcast content effectively, the solutions to these challenges are different for each individual. These differences mean that any effective technological solution must be personalisable.

The perceptual studies undertaken in Part II have evidenced some of these differences. They have demonstrated that redundant, non-speech sounds can aid broadcast speech perception in normal hearing individuals. Furthermore, they have shown that this benefit is extended to some, but not all, hard of hearing listeners.

Part III of this work aims to develop a technological solution which allows users to adjust the quantity and loudness of non-speech sounds based on how useful these sounds are to them. In order to be viable in the broadcast chain, it must be a solution which is intuitive to both production staff as well as end-users. Furthermore, it must be respectful of the creative integrity of the production staff's work. To achieve this requires a way of prioritising and thinking about sounds which captures the meaning and value of the sounds for both content creators and end-users. Only by establishing this common language can technological solutions which work for both production staff and end-users be created.

[1] The terminology adopted in this work for referring to narrative importance is to use the full term to refer to the concept of narrative importance or a narrative importance approach. The abbreviation NI is used for technical implementations – such as NI metadata assignment, NI prototype and NI control.

Usefulness of audio objects

Television constructs meaning through narrative. In Cohen's words, in television *'[...] narrative is primary, and the audience member is actively engaged in constructing a narrative'* [355]. This underlines the fact that each individual viewer constructs their own version of the narrative from what they see and hear, as well as from the context in which they observe these elements [21]. Previous sections of this work have focused on the usefulness of dialogue for conveying meaning and non-speech sounds in supporting that dialogue. However this is far from the only manner in which audio objects, both speech and non-speech, function to support a television narrative. This section gives a short overview of the different roles audio objects can play in television content.

Most works delineate television audio into four categories: dialogue, music, sound effects, and ambiences/background sounds [356, 21, 170]. Butler argues that the primary function of television sound is to direct attention as, unlike film, television is usually viewed in an environment like the home where competing stimuli are present [357]. The roar of a football crowd for a goal or a gunshot in a murder mystery are sounds which will draw back a viewer whose attention has been lost.

Deutsch [356] argues that these four groups can be further categorised into: *literal sounds* whose role it is to encourage us to believe what we see and *emotive sounds* which encourage us to feel something about what we are seeing. Cohen [358] suggests that music alone can take on the role of:

1. Masking – of unwanted noises or gaps on the soundtrack.
2. Provision of continuity – between scenes or locations.
3. Direction of attention
4. Mood induction – to express a redundant or contrary emotion to the on screen action.
5. Communication of meaning
6. As a cue for memory – through linking visual images with particular pieces of music, the music can act as a reminder of a particular event or trigger for a particular emotion.
7. Arousal and focal attention – through activating portions of the brain associated with music.
8. Musical aesthetics – increasing enjoyment of the content through enjoyment of the music.

Whilst Cohen limits her discussion to music, the majority of audio objects within a soundtrack can be utilised in the same way for the above purposes. Sound effects can be used to establish location, as well as on or off-screen action and they can also be used to

convey the narrative of off-screen events or for comedic effect [21]. These roles are not limited to dramatic content either in that sound effects can also draw attention in sports, like a referee's whistle, or construct aspects of the narrative, like the ricochet of a ball off the net in an attempted goal [359, 360] .

It has been said that if one were to *'collect all the studies and articles printed on the various effects of television'* one would probably need *'to hire a big truck to haul it all away'* [361]. This is similarly true of the effects of the sounds in television soundtracks. So with such a rich array of sounds, how can they be prioritised for listeners to ensure the narrative is conveyed without overloading their auditory or cognitive capacity?

9.2.2 Prioritisation in audio description

In this endeavour, inspiration can be taken from the established accessibility domain of Audio Description (video description services in the US). Audio description is an access service provided for blind and partially sighted individuals, whereby an additional descriptive commentary is interwoven between existing dialogue and critical sound effects [362]. This additional commentary provides a verbal picture of the scene for those unable to perceive the visual image themselves [363]. Like soundtracks, the visual elements of television present more things than one could ever possibly describe. Even if the time was available to describe every scene in exact detail, doing so would overwhelm the user with superfluous detail. For this reason, the describer must make the difficult judgement of "what not to describe" [364].

Fryer [362] suggests that these judgements on priority can be resolved, or at least simplified, by considering Gibson's theory of affordances – what the elements of visual scene affords the viewer? To illustrate this theory she, like Gibson, asks the reader to consider a scene containing a cat sitting on a mat. What we see is *'the mat extending without interruption behind the cat, the far side of the cat, the cat hiding part of the mat, the edges of the cat, the cat being supported by the mat, or resting on it, the horizontal rigidity of the floor under the mat, and so on'* [365]. These are all physical qualities of the cat and the environment in which it exists. Detailed description of these physical qualities accurately represents the scene but does not aid the blind or partially sighted viewer in their interpretation of the narrative.

Instead, Fryer suggests the key question the audio describer should ask themselves is *'what does the cat on the mat afford?'*[362]. For example:

'If the cat is twitching its tail, it is less likely to afford us the opportunity of stroking it. Similarly if its coat is scabby or crawling with fleas. If the cat features in a film we are describing, it may afford information about its owner – the Persian cat stroked by the

Bond villain Blofeld in To Russia with Love, indicates Blofeld's interest in prestige and appearance.'

By considering what each element affords the viewer in their construction of the programme's narrative, prioritisation of the multifarious visual elements becomes a manageable task.

9.2.3 Gibson's theory of affordances

In Gibson's seminal work on the theory of visual perception, *'The Ecological Approach to Visual Perception'* [365], he puts forth the argument that objects and environments are not neutral . Objects and environments *affords* the animal something and these affordances differ from their physical qualities. Physical qualities perceivable through our visual modality include colour, texture, composition, size, and shape, among others. For example, a ball may have the physical qualities of being round, green, and rubber but what it affords us is the opportunity to throw, catch, and bounce. If that ball has the qualities of being large and made of stone, it no longer affords throwing (unless accompanied by a trebuchet). However, it may afford sitting or a hiding place.

These affordances themselves can be positive or negative, beneficial or injurious; though, as Gibson cautions, these *'are slippery terms that should only be used with great care'*. Consider a substance that affords ingestion. This substance might afford nutrition or it might afford poisioning. Crucially, what the object affords is not only a function of itself but also of the observer. A peanut might afford nutrition to me, but, to a observer with a peanut allergy, it might afford anaphylaxis. A steak dinner might afford the observer a delicious treat, but, if the observer is a vegetarian, it may afford feelings of disgust. These affordances are not intrinsic to the item but are *'taken with reference to the observer'* [365].

Perception of these affordances does not classify an object or to be more precise *'the fact that a stone is a missile does not imply that it cannot be other things as well'*[365], or in the words of Jake Peralta *'stuff can be two things'* [366]. The differences between objects are not clear cut and this is why uniquely specifying objects and the categorises in which they belong have proved so philosophically elusive. But an inability to classify and label an object does not preclude you from understanding its usefulness. Gibson states that if you know *'what it can be used for, you can call it whatever you please'*.

Affordances in audition

Gibson's work is limited by his focus on only the visual modalities of perception. Others have since extended his ecological approach from visual perception to audition. The first

to give considerable attention to the idea of *ecological acoustics* was Gaver [367, 368]. His theories put forth the idea that we primarily perceive sounds as *'events in the world'* rather than as bare combinations of spectro-temporal characteristics perceived in the brain. Audition is an environment-oriented affair which is used to inform observers of real-life objects and events occurring around them. This theoretical perspective is supported by experiments which demonstrate the human ability to detect the properties of objects, such as their length [369] and geometric form [370]

This step to ecologise the theory of audition takes the perception of sound from being a purely cognitive endeavour, to one which is rooted in our environment. However, Gaver's theories still consider the observer as a passive vessel, ready to be filled with information from the environment [371]. He does not adopt the core tenet of Gibson's theory of affordances that perception is a relational experience and not an invariant function of the environment, but a function of the interaction between the environment and the observer [365].

Work by Steenson and Rodger argues that in order to more usefully characterise sounds, we need to consider not only how their physical qualities relate to the environment but what they afford the listener. With the exception of music [372–374], affordances have received relatively little attention in the field of audition. Yet, the concept of this relational development of perception extends as naturally to audition as to visual perception. The sound of a car approaching affords avoidance, or alternatively, affords injury. Music affords movement, inviting *'a synchronous motor response'* [374].

This conceptualisation of auditory affordances is backed up by perceptual experiments. Perception of the size of an object from its auditory characteristics when dropped results in significant difference in the apertures of an individual's grip when asked to reach and grasp that object [375]. It has also been shown that vocal instruction, using a voice alarm system, affords greater urgency in evacuations than a simple alarm tone [376]. When unsuspecting experimental participants are exposed to either a fire alarm tone or voice alarm, pre-travel activity time (time elapsing before evacuation) was significantly shorter in the case of the voice alarm.

As in visual perception, we appear to be active agents in auditory perception, perceiving not only the physical qualities of our auditory environment but what these sounds afford us.

9.2.4 Introducing 'narrative importance'

'The theory of affordances rescues us from the philosophical muddle of assuming fixed classes of objects, each defined by its common features and then given a name.'

– Gibson [365]

The common classification of television sounds into four categories – dialogue, music, sound effects and ambiences – groups the audio objects by their physical qualities [170]. *Prima facie*, it is simple to determine what should be contained in the dialogue group; it should contain a collection of sounds with the spectro-temporal characteristics of speech, made up of units of speech sounds (phonemes) and ordered according to standard patterns of speech sounds in a particular language. In production, this classification by physical qualities can be useful for applying group level signal processing such as compression and equalisation to sounds with comparable characteristics. However, one does not need to ponder far to find examples which equally meet these physical qualities but play a very different role in the television soundscape; background chatter in a restaurant or an *a cappella* piece of music. These have spectro-temporal qualities more similar to dialogue than other sound categories, but do not play the role of dialogue.

If these sounds are considered instead based on what they afford, we are rescued from the philosophical muddle of fixed classes. With the cat on the mat somewhat weary of being referenced, let us instead consider the dog on the rug. To visually describe the scene, we see a brown greyhound sitting on a pink and purple rug in front of a tired looking brown couch. The dog has big, wide eyes and no collar. To acoustically describe it, the dog might be emitting a high pitched whine or repetitively clawing at the material on the couch. What this affords us is an understanding that the dog is distressed by the fact that his owner has left, maybe clawing at his owner's favourite spot on the couch. Should this change to chipper, rapid barks and the rhythmic thud of the dog's tail wagging against the couch, it affords us knowledge that the absent owner has returned.

As in the visual domain, we need not be able to describe the physical qualities of the sound to know what it affords us. Dialogue may afford us the emotion of the speaker through tone and prosody without requiring the viewer to know what the term prosody means while music might afford a cue for memory, without requiring classification of its genre or tempo. A particular listener need not be able to identify the number and gender of the speakers in the background chatter of a restaurant scene to perceive that it affords them distraction and masking.

It is evident that Gibson's theory of affordances is congruent with the conception of the role of sounds within a television programme but is also congruent with the active construction of narrative by the viewer described by Cohen [355]. Listeners are not passive; they will infer meaning based on their experiences and what they need from the programme and the context in which they are observing the content. This subjectivity is, in Gibson's terms, is *'a necessary and inevitable feature'* of television content.

This leads us back to the prioritisation approach for the audio describer, put forth by Fryer [362]. Not only is the affordance of each object for the viewer considered, but a hierarchy of importance is established. Let us call this prioritisation of sounds 'The hierarchy of narrative importance'. Conceptually this hierarchy already exists in the content creator's production process and to subsequently exploit this in the creation of object-based personalisation tools, we need only a means to capture this perceived hierarchy in metadata.

How then to account for the second observer, the programme viewer? By actively constructing narrative as the viewer we are, in essence, subjectively altering the content as we consume and interpret it. Furthermore, when consuming content we will take steps to optimise the content, to maximise information, and to minimise external masking by turning up the volume [15, 172], telling the children to be quiet, or putting on a set of headphones. Allowing control over the elements of the audio mix through content personalisation simply formalises this optimisation process.

The binary paradigm of speech and masking used in much of early accessible audio work [140, 25, 147–149, 152] can be considered a crude attempt at classifying the soundscape by what it affords the viewer. However, by basing itself on the assumption that all speech affords meaning and all non-speech sounds afford masking, they have robbed the viewer from taking an active role in constructing the affordances of the soundscape.

Instead, this work proposes that the end-user has access to this scale of objects. At one end of the scale, all objects are at the level the producer intended, the 'TV mix' if you will. Then, as you move down the scale, objects which are less important are attenuated and the level of more important objects is increased. This then, at all times, retains the sounds most crucial to conveying the creator's perception of the narrative whilst offering the end-user significantly more control to minimise maskers and optimise key elements.

9.3 System design

This hierarchy of sounds gives a single scale for audio objects which is interpretable to both content creators and end-users. This differs from previous approaches which either utilised categories based on the physical qualities of sounds [170] or binary paradigms [218, 140, 25, 147–149, 152]. The former, whilst well received, presents an overly-complex interface requiring the control of four different volume controls. Approaches based on a binary paradigm resolve this complex interface by having a single control but do so at the cost of the content's comprehension and integrity.

By establishing a single meaningful scale for the audio objects means that we can provide end-user personalisation of these objects with a single control without compromising the

content's integrity. The remainder of this section describes the proposed accessible audio system, using object-based personalisation, centred on this idea.

9.3.1 Quantisation of the scale

As we perceive the world in analogue, continuous values, it follows that the proposed hierarchy of sounds should occupy a continuous scale. However, encoding of a continuous scale in a system is not a practical endeavour and subsequently the scale must be quantised.

This work proposes a scale with four levels, giving it significantly more granularity than the binary approach but intended to remain easily manageable. These four levels are given monikers to aid their interpretation which are:

- 0 - ESSENTIAL
- 1 - HIGH IMPORTANCE
- 2 - MEDIUM IMPORTANCE
- 3 - LOW IMPORTANCE

This work proposes that for each audio object in an object-based production, one of these values be assigned and stored in a new metadata field termed 'Narrative Importance – NI'. This value is assigned by the content creator. The detail of the proposed process for acquiring this NI metadata, and evaluation of this process with production staff, is described in Chapter 10.

9.3.2 Gain laws

The quantised scale allows for the first observer's (or content creator's) perception of the affordances of the audio objects to be recorded. Within what range should the second observer (programme viewer) then be able to adjust the level of these objects? This can be built on the pragmatic results of Chapter 8 which suggested that for relevant non-speech sounds at -3dB (relative to the speech), masking sounds should be able to be reduced by up to 15dB for those with mild to moderate hearing loss. Referencing the range of gains of attenuations to objects in HIGH IMPORTANCE rather than speech, as in Chapter 8, allows for ESSENTIAL objects (likely speech) to increase in loudness with reference to their level in the original mix. To accommodate for the extended range of attenuation potentially required by those with higher degrees of hearing loss, this work proposes that the range of the LOW IMPORTANCE extends to -48dB with reference to the HIGH IMPORTANCE objects. This effectively allows sounds in this category to be reduced to a level where they are imperceptible. From this, the range variables for each category can be defined and presented in Table 9.1.

Table 9.1 Level multiplier values (in dB) used in Eqn. 9.1 for each narrative importance level.

Narrative Importance level *I*		**Multiplier value *M* (dB)**
0	ESSENTIAL	+3
1	HIGH IMPORTANCE	0
2	MEDIUM IMPORTANCE	-12
3	LOW IMPORTANCE	-48

To then link this to a single control, the audio objects in each of these categories could have their level varied linearly with reference to the control. If the control exists in a range from 0 and 1, the following equation can define it's behaviour:

$$\text{Level}_{\text{adj}}\,(\text{dB}) = \text{Level}_{\text{orig}}\,(\text{dB}) + \Big[M_I \times N\Big]\,(\text{dB}), \tag{9.1}$$

where M is a multiplier depending on the narrative importance level I, and N is a balance between the NARRATIVE and IMMERSIVE settings (ranging from 0 to 1 for fully IMMERSIVE and fully NARRATIVE respectively). The multiplier values for the four narrative importance levels are detailed in Table 9.1.

9.3.3 Prototype system

To trial the hierarchical narrative importance concept and the proposed end-user personalisation based on it, a prototype system was developed.[2] This prototype system required input of an object-based audio file with corresponding narrative importance (NI) metadata for each audio object.

System function

The first prototype system was based on the 'meta-adapter' framework for intelligent metadata adaptation described by Woodcock et al. [377] and Figure 9.1 shows the flow of audio and metadata through the prototype system.

The narrative importance for each audio object was encoded into an ancillary .json file for the prototype. It was formatted with the intention that it could be easily adapted into a Broadcast Wave Format (BWF) file with Audio Definition Model (ADM) ITU-R [379]

[2]This prototype system was developed collaboratively between Dr. Ben Shirley, Dr. Jon Francombe and the author, as part of the S3A: Spatial Audio for the Home project. Dr. Jon Francombe programmed the function and interface for this prototype.

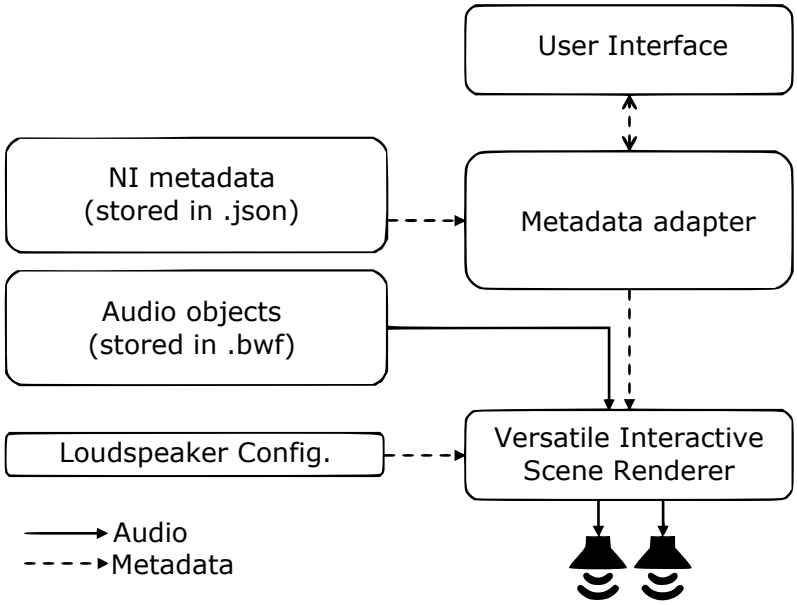


Figure 9.1 System diagram of the narrative importance control showing the flow of audio and metadata. Adapted from system diagram published in [378].

metadata. Metadata are passed to the *Metadapter*—a *Python* software package for metadata adaptation based on rulesets and user input [377].

The audio and adjusted metadata are then passed to the *Versatile Interactive Scene Renderer* (*VISR*) [380] which used amplitude panning [381] to generate loudspeaker feeds for the reproduction system described in a configuration file.

A simple single dial interface was developed in *Max/MSP* (Figure 9.2) with which the users could personalise the content. The dial could be controlled with a hardware rotary fader (*Griffin PowerMate USB*).

Content

For the prototype system, two pieces of object-based content had NI metadata added which were: 'The Turning Forest' and the 'Protest Scene' [382]. Metadata assignment was completed by the engineer who programmed the prototype.[3]

9.4 Study One: Initial user-experience evaluation

This section undertakes user-experience evaluation of the prototype system for personalising broadcast audio as developed in Section 9.3.3. A focus group methodology was selected as it

[3] Dr. Jon Francombe.

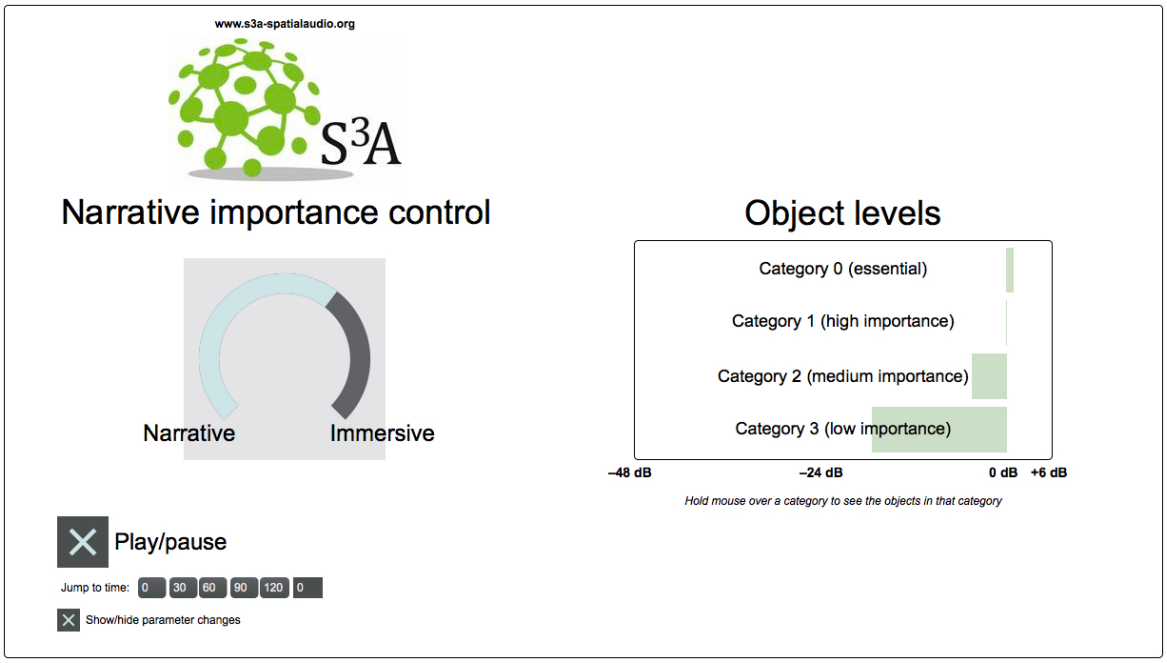


Figure 9.2 Interface of the narrative importance control prototype, showing the NARRATIVE to IMMERSIVE scale and the attenuation/gain for each narrative importance level.

allows for exploration of both user's experience of trialling the prototype, in depth discussions of what they feel works well and discussions of what could be improved. Furthermore the group structure of the method allowed for differing views to be expressed and individuals to be challenged in their views [383]. In particular, it allowed the experience and views to be compared and contrasted by those with varying degrees of hearing loss as well as those with and without hearing aids.

This section first describes the focus group methodology and how the focus group transcriptions were analysed. The results of the focus group are reported and finally the key conclusions from the user-experience evaluation are discussed.

9.4.1 Focus group methodology

The focus groups took place at the University of Salford, either in the listening room or a soft seating area in the Newton Building. Each focus group began with an introduction, giving the participants the context of the focus group and an outline of what the prototype control did and why it was developed. The introduction was scripted to ensure consistency between the focus groups. The introduction was as follows:

Thanks for taking the time to join us to talk about problems with TV sound. We would like to better understand your experience of TV sound and try out a tool that changes the mix of

the sound. We would like to know what you like, what you don't like, and how the tools might be improved. There are no wrong answers but rather differing points of view. Please feel free to share your point of view even if it differs from what others have said. Keep in mind that we're just as interested in negative comments as positive comments, and at times, the negative comments are the most helpful. We're recording the session because we don't want to miss any of your comments. People often say very helpful things in these discussions and we can't write fast enough to get them all down. We will be on a first name basis today and we won't use any names in our reports. You may be assured of complete confidentiality.

After the introduction, each participant took turns to trial the prototype whilst other participants listened. The prototype was run from a laptop with the audio reproduced over two Genelec 8030A Studio Monitors. The prototype was controlled via a Griffin Powermate USB Multimedia Control Knob. As this controller is continuously rotatable with no fixed endpoints, the user interface in Figure 9.2 giave the user a visual indication of when the endpoints of the scale are reached.

The focus group participated in both structured and unstructured discussion. Participants were encouraged to discuss any related topic they wished but, as recommended by Bryman [383], the facilitator used a small group of general questions to guide the focus group. These can be seen in Table 9.2. These general questions were used to catalyse discussion, particularly for participants who were unacquainted before the focus group, and maintain the flow of conversation throughout. The scripted list of questions can be seen in Table 9.2.

Table 9.2 General questions used by facilitator to guide the focus group discussions.

	Questions
1	What did you think of the control?
2	Do you feel this makes the programme clearer?
3	What made you set the level as you did?
4	What elements of the sound were useful?
5	Think back to when you have watched TV at home, do you ever have trouble understanding what is said on TV?
6	If so, what do you think caused the problem?
7	Would you use the control at home if it was available?
8	What would you change about the control to make it better?

Participants

A targeted cohort of older and hard of hearing listeners were recruited for the focus groups. All participants were either over the age of 50 or identified as having mild to moderate

hearing loss. Focus groups were conducted at the University of Salford as part of the same event where data for Study Four in Chapter 8 was collected.

Sixteen participants took part in 4 focus groups with 3 to 5 participants in each. In addition, two individual participants took part in semi-structured interviews, giving 18 participants in total. Focus groups lasted between 20 minutes and 1 hour depending on the size of the group.

Participants' pure-tone audiometric averages (0.5-4kHz)) were 35 dBHL in their better hearing ear and 47 dBHL in the worse hearing ear. Hearing loss included bilateral, unilateral, and high frequency age related hearing loss. Participants' ages were between 53 and 95, with a median age of 64. Eight participants had had hearing aids fitted but only half of these identified as wearing them regularly.

All focus groups were facilitated by Dr Ben Shirley and audio recorded using an Olympus DM-1 voice recorder. The majority of focus groups also had an additional researcher in attendance to help facilitate discussion and note down key points.[4] At their completion, all focus group recordings were transcribed either by the author or by an external transcription company.

9.4.2 Analysis methodology

To analyse the participant responses and discussions, content analysis was selected as in Chapter 3. This approach was deemed more suitable than Grounded Theory as Grounded Theory assumes no existing hypothesis and develops any hypotheses directly from the data. Inductive content analysis is more suitable here as this focus group was testing a specific hypothesis, that the developed prototype improves the experience of television sound for hard of hearing individuals. Furthermore, as a particular hypothesis is being explored, the analysis will focus on themes rather than patterns of interaction.

Identification of the key themes was performed iteratively by two raters, the author and the focus group facilitator. Only comments which pertained to the prototype were considered, as many tangential topics arose in the focus group discussions.[5] One rater developed the initial framework through open coding of the focus group text, guided by the questions in 9.2. From this a framework with five high level themes: Interface, Agency, Content, Speech and Balance, with 3-5 sub themes each was developed. Coding was then performed independently by the second rater, utilising this framework. The second rater made additions and deletions

[4]For all bar one focus group, the additional researcher was Philippa Demonte. The author collected notes at the final focus group.

[5]Including a treatise on the need for a technology which can slow down hip-hop and make it more accessible. This can be found in full in Appendix F.

to the coding framework, based on the quotes identified. Then, through discussion between the raters, a final coding scheme was developed, which combined Content and Speech into a single category and altered the hierarchy of themes in Interface. The final coding framework can be seen in Table 9.3.

Table 9.3 Coding framework developed from the focus group text data using Inductive Content Analysis.

High Level			
1. Current Prototype	**2.** Agency	**3.** Balance	**4.** Content
Low Level			
1.1 Positives **1.1.1** – Would you use it at home? **1.2** Negatives **1.3** Features **1.3.1** – Compatibility **1.3.2** – Profiles **1.4** How did you set it?	**2.1** Editorial Decision **2.2** Single/multiple users **2.3** Stigma **2.4** Cost	**3.1** Greater Range **3.2** Non-speech behaviour **3.3** Separate control of objects **3.4** Role of non-speech objects	**4.1** Visuals **4.2** Speech and Tone **4.3** Adverts **4.4** Genre

All identified quotes were sequentially addressed by the two raters and, where they were coded differently, were discussed and resolved to a single code. This resulted in 150 individual quotes being identified. This approach differs from that used in Chapter 3, where inter-rater agreement was calculated. Given that the framework was being iteratively developed in tandem with the coding process, a discussion based resolution process was considered more appropriate than calculation of inter-rater agreement.

9.4.3 Results

To visualise the key points from each of the high level themes, results were aggregated and displayed in Figures 9.3, 9.4, 9.5 and 9.6. In each diagram, the number of codes in that category are indicated by the numbers adjacent to the lines linking the low level theme to the high level central theme as well as the weight of the line. For each of the low level themes, 1-3 quotes exemplifying the key topics related to that theme are included. These key themes and quotes are discussed in the following sections.

Significance testing was then undertaken using a Chi Squared Goodness of Fit test to determine first if the codes were evenly distributed and, if not, which codes occurred

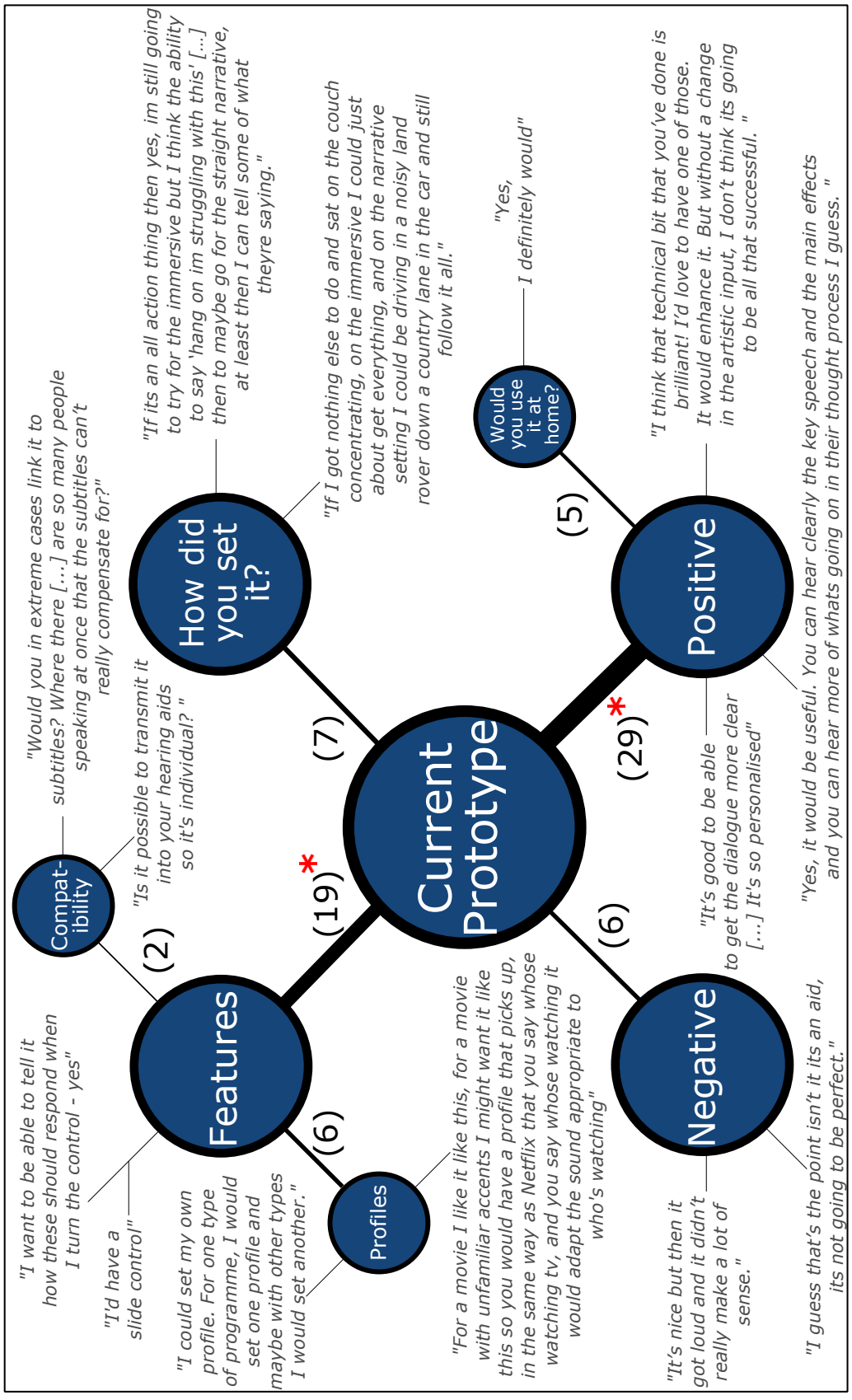


Figure 9.3 Visualisation of low level themes identified in the focus group data relating to the high level theme: **Interface**, with counts for each category and example quotes noted. Categories which were identified significantly more often are noted with a red asterisk.

significantly more than others. Whilst this indicates which categories of responses should be given greater weight in the analysis, all responses are considered important in evaluating the prototype's function and are included in the discussion in the following section. The Goodness of Fit test yielded a test statistic $[z = 98.1]$ which was significant at the level $[p \leq 0.01]$, for $df = 15$. The upper and lower bounds of significance were calculated as $C_{sig} = 15.4$ and $C_{sig} = 3.4$ respectively, based on a two tailed test at the level $\alpha = 0.05$. Three codes were found to occur significantly more often: 1.1 POSITIVE – 29, 3.3 SEPARATE CONTROL OF OBJECTS – 23 and 1.3 FEATURES – 19 . A further three codes were found to occur significantly less often than others: 2.4 COST – 3, 2.3 STIGMA – 2 and 4.3 ADVERTS – 1.

9.4.4 Discussion

The prototype was very positively received by the focus group participants (Figure 9.3, with positive comments about the interface being the largest single code identified (29). When asked, all participants agreed that with the NI control they were able to improve the sound and speech clarity of the content. One participant noted that they felt by reducing the LOW IMPORTANCE sound they could *'hear more of what was going on in their [the post-production staff's] thought processes'*. Though, as one participant notes, the technology's effectiveness is reliant on the quality of the content: *'without a change in the artistic input, I don't think it's going to be all that successful'*. Responses were not unanimously positive with one participant indicating that generally they thought the prototype was nice but when the content got loud, the speech still *'didn't really make a lot of sense'*. On this note, one participant remarked *'it's an aid, it's not going to be perfect'*.

The second most commonly identified code was the SEPARATE CONTROL OF OBJECTS seen in Figure 9.4. Whilst participants appreciated the control that the system gave them over the content, participants differed strongly in their opinions on the value of the non-speech sounds and the degree to which the system should allow the viewer to control these objects. This reflects the results of Chapter 8 and is possibly related to the differing benefit these individuals gained from the non-speech sounds.

At one end of the spectrum, some individuals desired only speech and wished to turn off even the sound effects that might be important to the plot (Figure 9.4). They felt that *'If you lose the intonation [of the speech] the atmosphere is not very important, irrelevant'*. At the other end of the spectrum, participants felt that non-speech sounds were very important, *'because otherwise you're just listening to somebody reading a story'*. In particular, some felt that it helped their *'imagination of what it is I'm listening to'* and gave *'the depth and the fullness of what the guys in the studios have produced'*. When asked which sounds

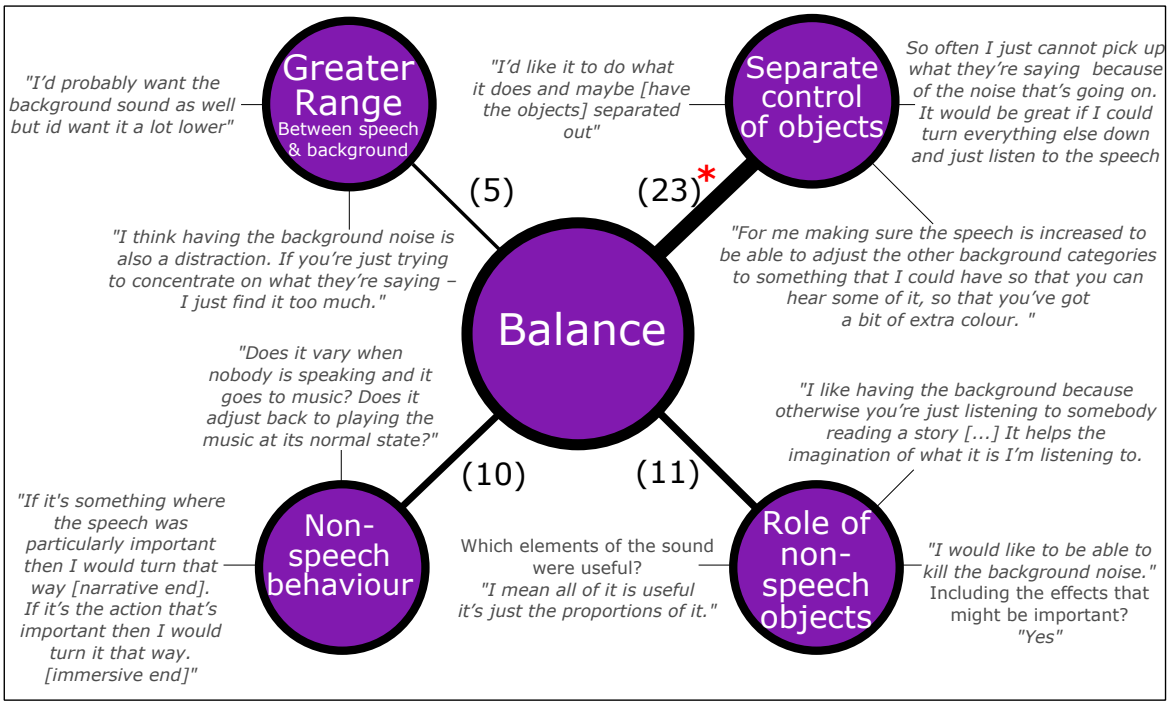


Figure 9.4 Visualisation of low level themes identified in the focus group data relating to the high level theme: **Balance**, with counts for each category and example quotes noted. Categories which were identified significantly more often are noted with a red asterisk.

were important, one participant responded: *'all of it is useful, it's just the proportions of it'*. Participants were also curious about whether the system behaved in the same manner when speech was not present, for example when only music was playing: *'Does it adjust back to playing the music at its normal state?'*.

For these reasons, many participants indicated that they would like to have separate control of the different objects, so that they could turn up the speech as much as they need to then add in some of the background, *'so that you've got a bit of extra colour'*. Other participants indicated that they would like the ability to turn the background sound down much lower than possible on the current prototype and that, even at the narrative end, they *'just find it too much'*. Participants also mentioned that the ability to change not just the volume but *'the tone of the speech, make it a little bit crisper as well'* would be advantageous.

FEATURES was the third of the most commonly identified code. Of these comments, the single feature which attracted the most discussion was some variation on a user profile. Participants desired this profile to adapt either to different people who might use the same television (in the manner of Netflix's 'who's watching?': see Figure 11.1) or profiles for different genres of content (*'For one type of programme, I would set one profile'*). The desire to adapt to different kinds of content is motivated by the different ways individuals envisaged

setting the control for different genres (see: Figure 9.5). Some participants indicated there were genres where they didn't feel they would use the control (e.g. sport and news) and genres where they would (e.g. drama and film). Other individuals indicated they would probably use it all the time, but want different settings for the different genres; *'a bit more of the background sound'* for a TV drama, but *'a bit less'* for a blockbuster film. These discussions highlighted that genre-based profiles would still need to be individualised to the listener, rather than being a generic profile for a particular genre.

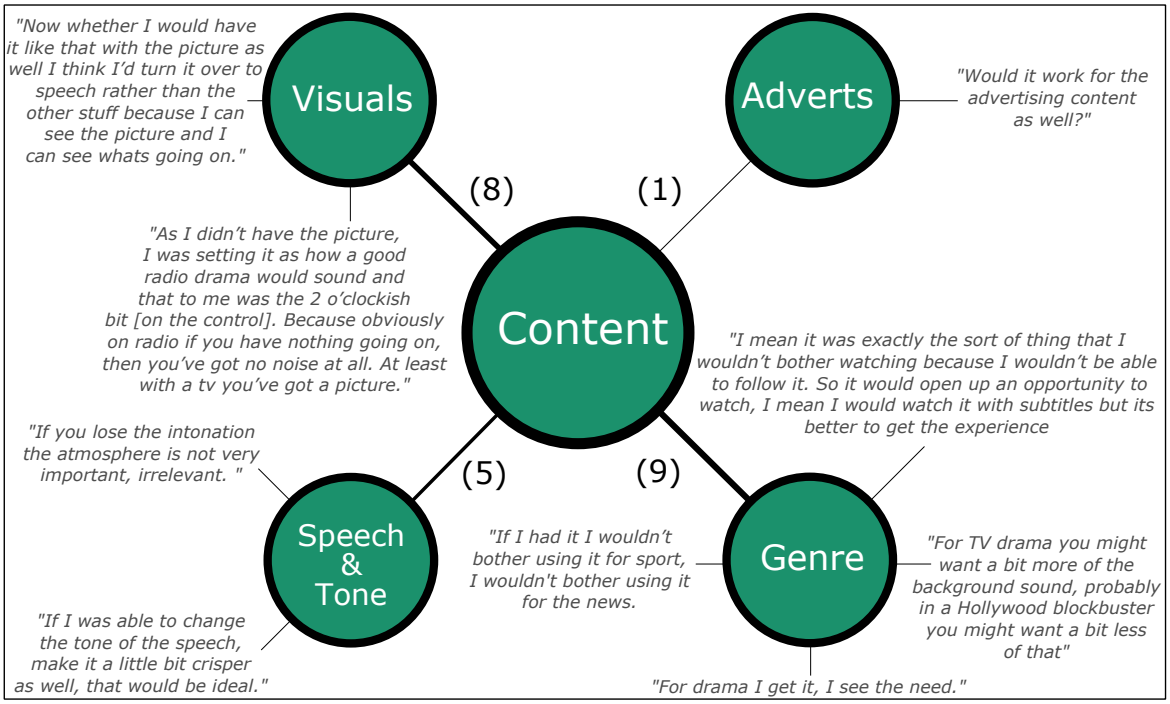


Figure 9.5 Visualisation of low level themes identified in the focus group data relating to the high level theme: **Content**, with counts for each category and example quotes noted.

The use of profiles for different listening scenarios was highlighted by one participant who envisaged that their setting would be dependent on where they were listening as well, comparing *'sat on the couch concentrating, on the* IMMERSIVE *[setting] I could just about get everything'* with the NARRATIVE setting where they could be *'driving in a noisy Land Rover down a country lane in the car and still follow it all'*. Beyond profiles, participants were interested to know whether this technology would be compatible or linked with existing systems, like subtitling and their hearing aids. In particular, one participant asked whether it could provide clearer speech in scenarios where subtitles currently performed poorly, when there *'are so many people speaking at once that the subtitles can't really compensate for?'*. Participants also discussed current aspects of the prototype, like the interface (which occupied the whole laptop screen) and the rotary knob, and whether these would be the same in future

iterations of the technology. Participants indicated that a control integrated into the remote or set-top box, or a slider control, would improve the experience of using the technology.

In setting the control, many participants noted that they would likely set it differently if the content had visuals. The reasons they identified included that radio drama is busier as the sound is the only way of conveying narrative: *'Because obviously on radio if you have nothing going on, then you've got no noise at all.'* In contrast, another participant indicated that they would set it more to the NARRATIVE end for audiovisual content: *'I'd turn it over to speech rather than the other stuff because I can see the picture and I can see what's going on'*.

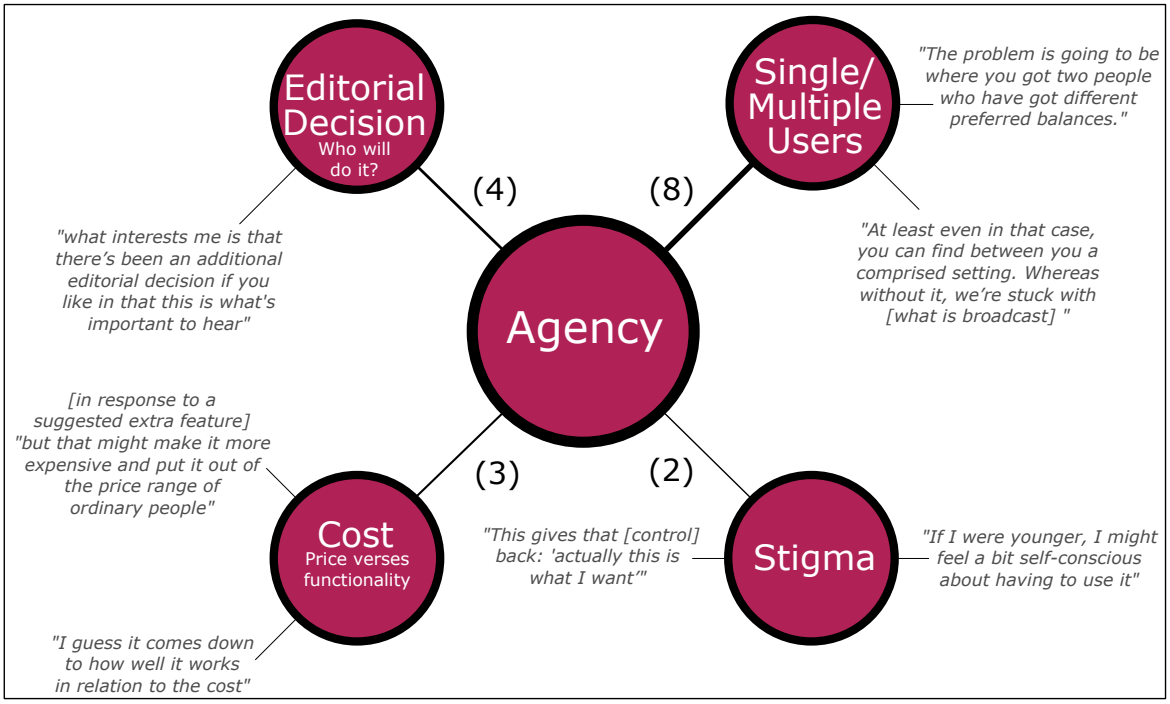


Figure 9.6 Visualisation of low level themes identified in the focus group data relating to the high level theme: **Agency**, with counts for each category and example quotes noted.

Many participants noted that by making the content personalised, viewing with multiple people might become a challenge: what happens when *'you got [sic] two people who have got different preferred balances?'* The facilitator discussed technology which could, through beam-forming and sound bar technology, allow different individuals to listen at the same time (without headphones) and receive different mixes [230]. Another participant also remarked that even though this presents a new challenge, you can *'find between you a compromised [sic] setting. Whereas without it, we're stuck with it'* (it referring to the current broadcast content). Linked to this discussion was a comment made about compatibility of

the technology where a participant asked whether the mix would be individualised by being sent straight to their hearing aids.

A common theme running through a number of categories was the concern that, if implemented in the standard broadcast system, the control might be exploited by producers and broadcasters (see: Figures 9.6 and 9.5). This included discussions relating to the editorial decision making that is involved in setting the NI metadata and that some content could be made louder or less controllable. One participant noted that in particular the technology could be exploited by advertisers who would make their ads unable to be adjusted.

Participants expressed interest in knowing when the technology would be available and at what cost. One focus group particularly dwelt on the idea that a control which was simpler but also cheaper would be preferable to one which was more expensive but which contained many of the additional features discussed. The general hope was that it would be affordable, in order to be available to as many people as possible. When queried about whether they felt they would use one, should they have it in their homes, they all agreed that they would.

Only two participants mentioned how the perceived stigma of deafness might interact with the prototype. One commented that needing the device might be seen as a negative: *'If I were younger, I might feel a bit self-conscious about having to use it'*. In contrast, the other participant remarked that the stigma of having hearing loss mean some people would not talk to them and the technology would give them back some control over their scenario: the ability to say *'actually this is what I want'*.

On this note, participants particularly liked the control that the technology gave them over their experiences. The importance of agency was echoed in how participant's described setting the control: *'I'm still going to try for the* IMMERSIVE *[end] but I think the ability to say 'hang on I'm struggling with this' then to maybe go for the straight* NARRATIVE*'* allows them to ensure that they get the story which was being conveyed. It was also highlighted that this would open up content from which they previously felt excluded. Describing 'The Turning Forest' content, one participant said: *'I mean it was exactly the sort of thing that I wouldn't bother watching because I wouldn't be able to follow it. So it [the control] would open up an opportunity to watch'*. This was also reflected in the desire for greater control over individual audio objects (see: Figure 9.4) as well as in the features they wanted: *'I want to be able to tell it how these [the different attentuations] should respond when I turn the control'*. The essence of the importance of agency was epitomised by one participant, who remarked:

'Being hard of hearing isn't a lot of fun but [this technology] puts us back in control. We are in charge of what we hear and I think that's quite empowering.'

9.4.5 Conclusions from Study One

It is evident from the focus group results that the prototype was thought of as a positive step for broadcast accessibility. The key aspects important to participants were firstly the control it gave them over their experience, and secondly, that they could balance the audio based on numerous personal factors: their attention, interest, listening environment and hearing needs. The majority of participants liked that they could control not just speech level but the amount of background sound to give the content colour and context. The variety of opinions on the value of non-speech sounds reaffirms the results of Part II of this work and highlight that the personalised approach is likely to benefit the widest range of audiences.

The majority of participants agreed that the ability to turn the speech up louder would improve the prototype. This could be implemented by restricting the essential category to only speech items. This would, under the current system, ensure that speech could be turned up at least 3dB above all other sounds and, under different systems, potentially more. Participants also wanted to have individual control of the object groups. However, by giving more specific control over the different objects is contrary to the aim of simplicity in the design of the system. It is also contrary to a recurring theme in the focus groups of wanting a low barrier to accessing the tool.

The focus groups identified numerous future features to investigate, the most popular of which was the creation of individual profiles. Additionally, the ability to change the background depending on whether there was speech present and to change the quality and tone of the speech were suggested as future features.

In addition to desiring the widest availability of the tool as possible, participants emphasised that they wanted to ensure that the additional control given to viewers was not exploited or restricted by content producers. This presents an interesting challenge as the same freedom that the system gives producers to preserve the creative integrity of their content, gives rise to the concerns the participants voiced including exploitation by advertisers. There is no easy solution to this challenge.

Finally, the focus groups highlight that, to properly evaluate the function of the system, more ecologically valid content which included visual elements would be needed. Furthermore, a more ecologically valid prototype, with greater similarity to current control mechanisms for media content such as remote controls, would also be required.

9.5 Chapter Summary

This section has described the development of a scale for audio objects which allows for the value they afford both the content creator and the end-user to be encoded. Based on this

novel approach to prioritising sounds in television content, an accessible audio system using object-based personalisation methods was proposed. In addition to defining the function of the proposed system, a prototype of the system with two pieces of object-based radio drama content was developed.

This prototype then enabled initial user-experience testing of the proposed system to be undertaken with a group of target users. A focus group methodology was selected for this evaluation as it allowed for discussion about the functionality of the interface as well as broader discussion of the concepts behind the prototype. The key results from the focus group were:

- Participants felt that the system would improve their experience of broadcast audio and appreciated the agency over their media experience.
- Participants differed in their opinion of the importance of non-speech sounds but the majority like the additional *'colour'* and *'depth'* the non-speech sounds gave.
- Participants wanted to be able to turn speech up louder (or turn background sounds down further) and the ability to balance the audio to their own preference.
- Profiles, to simplify regular use of the system, were identified as a key feature that the participants would like to see included in the system.

These results suggest that a system which allows end-users to control the balance of audio objects whilst preserving comprehension is desirable for end-users. However, the viability of this approach needs to be evaluated with production staff and thus forms the focus of Chapter 10. The results also indicate that further investigation of the merit of the proposed system is required, in particular in a more ecologically valid implementation and with audiovisual content. This investigation is described in Chapter 11.

Chapter 10

Production Workflows for Narrative Importance

10.1 Introduction

Chapter 9 demonstrated that it is feasible to design a system that *allows end-users to control the balance between audio objects for dramatic content which is simple to use and preserves comprehension*. The proposed system intends to capture the content creator's conceptualisation of what audio objects afford the narrative and the viewer. Furthermore, through object-based metadata, it aims to encode how content creators prioritise audio objects within a television soundscape.

To determine whether the proposed approach achieves these aims, there are two key questions which must be addressed:

- Is the concept of narrative importance commensurate with production staff's own prioritisation strategies for audio objects?
  - And does this conceptualisation differ between production staff?
- Is it feasible to integrate narrative importance (NI) metadata [1] acquisition into current production workflows?

Study Two begins to answer these questions using qualitative research methodologies and user-experience evaluations with production staff. This chapter begins with an in-depth ethnographic case study with the sound designer for two pieces of object-based audio drama,

[1] The terminology adopted in this work for refering to narrative importance is to use the full term to refer to the concept of narrative importance or a narrative importance approach. The abbreviation NI is used for technical implementations – such as NI metadata assignment, NI prototype and NI control.

'The Turning Forest' and 'The Protest Scene' [382]. This case study also involves trialling a Virtual Studio Technology (VST) plug-in for NI metadata assignment.

Study three explores how the results of this case study generalise to a broader community of production staff. This is achieved using an online survey and NI metadata assignment task with an excerpt from 'The Turning Forest'.

These two studies are then discussed indicating that the underlying concept of narrative importance is congruent with production staffs' processes and conceptualisation of audio object importance. Key functionalities required in future NI metadata assignment tools are also discussed.

10.2 Study Two: Case study with a sound designer

An ethnographic case study with a single sound designer was selected as the initial methodology for investigating production staffs' conceptualisation of narrative importance. This was selected as it allowed for in-depth exploration of the challenges NI metadata acquisition might present to workflows in a real-life production environment. Whilst only representative of an individual's experience and opinion, this was considered a valuable starting point on which subsequent investigations with a broader community of production staff could build (see Section 10.3).

This study was based around a two day production workshop coordinated with a sound designer. Due to the prototype nature of the technology being trialled, the author played an active role in the study guiding its direction.

This section first describes the methodology of the ethnographic study detailing the structure of the workshop, tools used and the recruitment of the participating sound designer. The results of the case study are then described, outlining the NI metadata assignment workflow developed by the sound designer and which narrative importance level was assigned to each piece of content's audio objects. A discussion of these results is presented, highlighting key areas for further investigation with a broader cohort of producers.

10.2.1 Methodology

Recruitment

The sound designer who produced the content used in Chapter 9's end-user prototype evaluations was approached to participate in the study. Upon agreeing to participate, two workshop days were scheduled where she would visit BBC R&D and work to generate the NI metadata for the 'The Turning Forest' and 'The Protest Scene' [382].

Before the first workshop, a preliminary phone discussion introduced the sound designer to the planned structure of the workshop and gave a brief overview of the concepts which the workshop aimed to explore. The workshop length was not fixed and was reflective of the time the sound designer took to complete the task. The sound designer was paid for her time.

Content – Synopsis

The content used for this case study were the two object-based radio drama scenes utilised in Chapter 9. This selection was motivated by the availability of the original sound designer to participate in the case study and the limited options for open-source object-based pieces of content available. The plot descriptions which follow are reproduced from [382].

'The Turning Forest' is a piece of content, 4 minutes and 40 seconds in duration, which describes a fantasy scenario set in a forest. *'The scene opens with two children playing in an autumnal forest. After one of the children runs away, the remaining child encounters a large friendly monster. The child walks with the creature through the forest and rides on the creature's back as it swims across a river. When the creature and the child reach the other side of the river, the season has changed from autumn to winter. The scene includes a narrator throughout and non-diegetic music.'*

The 'Protest' scene depicts a protest *'being staged outside a bank. The scene begins inside the bank and evolves from a front dominant image to full 3D sound as the action moves from indoors to the protest outside. The scene demonstrates immersive crowd atmos [sic], individually identifiable voices popping out of the atmos, and moving localisable sources at different heights.'* It is 2 minutes, 38 seconds in duration.

Content – pre-processing

Both pieces of content were originally produced for reproduction on a 22.2 loudspeaker layout [382]. As the intended application for the current work was television, which still primarily broadcasts in stereo, a stereo version of the content was more appropriate for the case study. Furthermore, use of a stereo down-mix prevented most spatial release from masking in the original influencing the NI metadata assignment. The original object-based mixes were rendered to stereo, with the positioning rendered appropriately.

The original mix of the audio dramas was completed in the digital audio workstation Nuendo [384]. To have the desired functionality in the metadata acquisition tool and interoperability with the other parts of the NI prototype, a new version of the mix was created in Reaper [385].

Prior to the workshop, the author decomposed the audio objects in the 'Protest' scene and 'The Turning Forest' into temporal 'conceptual' objects. For example, rather than a water splash being composed of multiple layered water sound effect tracks, they were rendered to a single 'water splash' object representing the intended conceptual audio object. These were verified with the sound designer at the beginning of the workshop to ensure they represented the original conceptualisation of the drama.

This task proved easier for the speech and spot effect objects than for the ambient and reverb tracks, particularly as the reverb tracks contained the reverb for multiple different objects. To circumvent this issue, the reverb tracks were not used for 'The Turning Forest' and a reduced set of three reverb tracks representing the majority of the left, right, and central content respectively was used for the 'Protest' scene. In the 'Protest' scene, the 9 channel spatial recordings of different types of crowd and ambience were also reduced down to left, right and central channels only.

All of the stereo audio objects in the two audio dramas were then rendered to individual left and right tracks for all objects for compatibility with the NI metadata plug-in which, at the time of the workshop, could only process mono objects (described in Section 10.2.1).

VST plug-in for metadata assignment

To acquire NI metadata, a VST plug-in for Reaper [385] was developed based on the suite of VISR plug-ins [386]. A screenshot of the plug-in can be seen in Figure 10.1. [2]

Each separate track in Reaper was treated as a separate object for which a name could be given. It could then have its narrative importance level assigned and be allocated to a group: groups could be created by the sound designer and given names as well as a narrative importance level. Metadata changes could be auditioned by the sound designer by sending the metadata to the NI control interface. This was also connected to the USB Griffin rotary fader, allowing the sound designer to alter the end-user setting on the narrative importance scale (as in the prototype in Chapter 9). The flow of audio and metadata can be seen in Figure 10.2. The audio was reproduced in stereo over two Genelec 8030B loudspeakers set at $\pm 30^{\circ}$ with reference to listening position of the sound designer. This was conducted in the BBC R&D listening room where the final 3D mix for the audio dramas was undertaken [387].

[2] This plug-in was developed as part of work on the S3A:Spatial Audio for the Home project. The plug-in was coded by Dr. Rick Hughes and supported by Dr. James Woodcock and Dr. Jon Francombe.

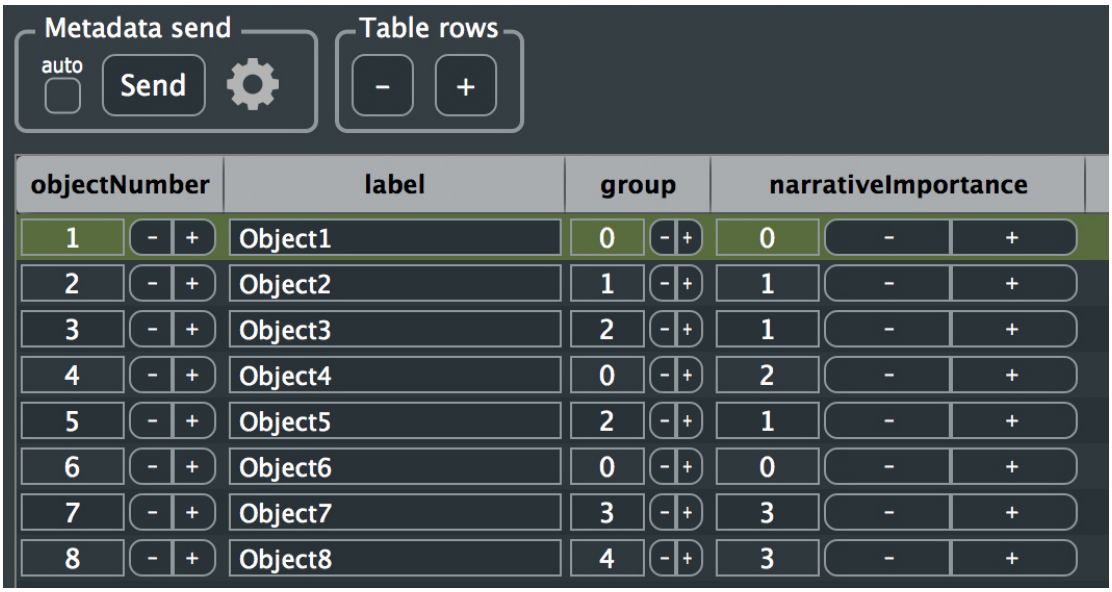


Figure 10.1 Screenshot of the NI metadata assignment plug-in, based on the Versatile Interactive Scene Renderer framework.

Workshop structure

The sound designer was asked to set the NI metadata for stereo versions of 'The Turning Forest' and the 'Protest' scene. They were given as much time as needed to complete the task.

The workshop began by introducing the sound designer to the VST plug-in which had been developed for the workshop and the NI end-user interface (developed in Chapter 9). She was then asked to verify the new stereo version of the mix in Reaper, broken down into stereo objects. The original Nuendo sessions were available for reference if needed (though these were not used).

The workshop was facilitated by the author with the sound designer encouraged to develop her own workflow for authoring the metadata using the plug-in. The NI metadata was first completed for 'The Turning Forest' and then the 'Protest' scene. Throughout the course of developing the workflow, the sound designer was asked to explain her rationale for the decisions made. The workshops were recorded using an Olympus DM-1 voice recorder and notes were also recorded using pen and paper by the author.

10.2.2 Results

The study took approximately 8 hours in total spread over the 2 days. 'The Turning Forest' was completed first and took the majority of this time; some of this time was also occupied

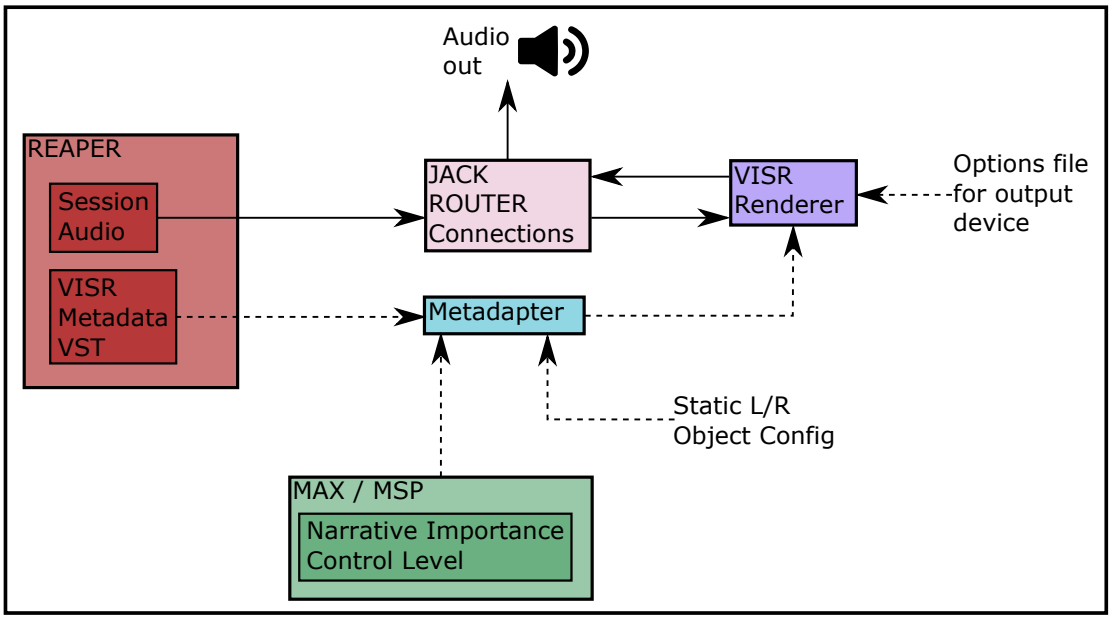


Figure 10.2 Flow diagram of the system used for Study Two, based on the end-user prototype described in Chapter 9 and the NI metadata assignment plug in shown in Figure 10.1.

by debugging small issues with the plug-in. The voice recordings of the workshop were transcribed by the author and are used to explain the described workflow and rationale for decisions in the sound designers' own words.

This section first reports the narrative importance level each of the audio objects were assigned to. How the sound designer conceptualised narrative importance and how she thought it could be integrated into current workflows are summarised then discussed in the context of the other results.

Metadata assignment results

The final metadata assignment for the 44 objects in 'The Turning Forest' and 47 audio objects in the 'Protest' scene can be seen in the Figures 10.3 and 10.4 respectively. For both pieces of content, only two objects were assigned to the ESSENTIAL group. These objects were either dialogue or effort noise, for 'The Turning Forest'. It can be seen that of the 44 audio objects in 'The Turning Forest', the majority are assigned to the lower levels. For the 'Protest' scene, the distribution of objects across the HIGH, MEDIUM and LOW IMPORTANCE categories is more even. This may be indicative of the type of audio objects in the 'Protest' scene which were primarily different kinds of speech.

The groups used in 'The Turning Forest' are noted in Figure 10.3 by coloured circles. The name given to each group is also noted in the corresponding coloured text. Groups

were not utilised as extensively in the 'Protest' scene and, as such, Figure 10.4 shows only individual object assignments. Colours denote similar objects, or versions of the same object, to aid interpretations of the diagram.

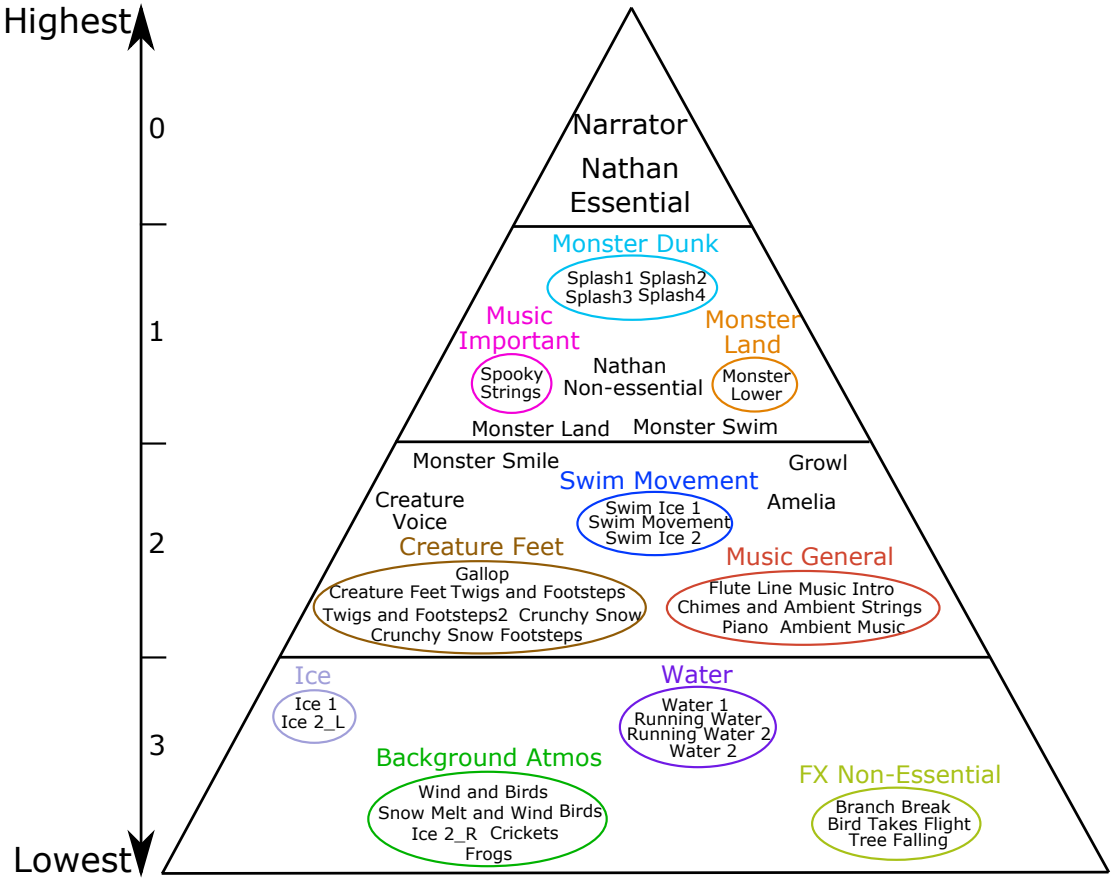


Figure 10.3 Hierarchical visualisation of the NI metadata assigned in Study Two to each audio object in 'The Turning Forest' with the groups set by the sound designer denoted by coloured circles and the group titles noted in corresponding coloured text.

Workflow development

The strategy developed by the designer for 'The Turning Forest' was to first name objects in the NI metadata plug-in and assign narrative importance level and groups. This was done by listening to the object individually and judging conceptually how important she felt it was in the mix.

The mix was then auditioned at both extremes— NARRATIVE and IMMERSIVE, with *'a bit of listening in the middle to see at what point it makes a massive difference.'* This approach was similar to one the designer regularly uses to validate that mixes translate from good quality stereo speakers to poor quality mono speakers: *'I make sure that the important*

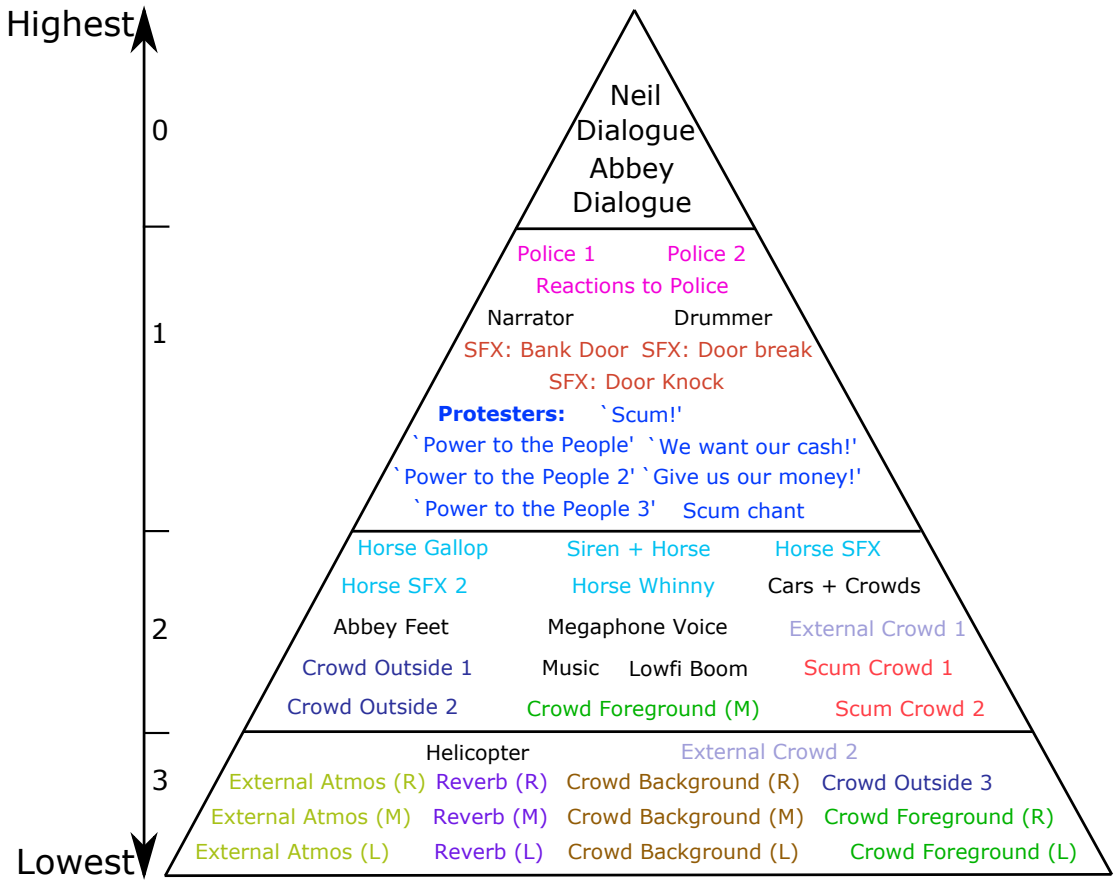


Figure 10.4 Hierarchical visualisation of the NI metadata assigned in Study Two to each audio object in the 'Protest' scene with colouring text denoting related audio objects.

speech is coming through and sometimes crucial effects but more often than not it's about the speech peeking through.' Expressing surprise that the difference between end-points of the scale on the end-user control were not as stark as she expected, the comment was made: *'it's actually quite subtle, isn't it?'*

Initially the designer found it difficult to allocate any objects to LOW IMPORTANCE: *'It's really funny, as a sound designer it's really hard to press three.* However after auditioning the mix with the NI end-user control, the designer reduced the importance of many objects whilst increasing the importance of a relatively small number of objects.

The designer experimented with the narrative importance levels for a number of longer objects containing multiple sounds of a similar type. In particular, the effort noises and emotive breathing of Nathan who is the protagonist in 'The Turning Forest', proved a challenge as the designer felt they conveyed important aspects of the story, but due to their long term nature also overlapped and masked some of the narration. Finally, she concluded that the object needed to be separated into multiple objects to effectively assign importance

'Nathan Essential' and 'Nathan Non-essential'. The main difference noted between the sounds in the 'essential' and 'non-essential' versions of the object was that those in the 'non-essential' temporally overlapped the narration. It was suggested that this could be addressed by having time-varying adjustment of the attenuation levels of categories 1–3. This would allow the levels of lower importance objects to swell when the most important objects were not present, ensuring that periods without narration or main characters were not too sparse.

Many of the decisions for the remaining objects were made by considering which sounds signposted a change in location or emotion, to *'give them a taste of where we are and then leave it.'* Similarly, sounds announcing the presence of the monster were set to higher importance than subsequent monster movement sounds and vocalisations: *'the monster lowering himself is more important than the feet that come afterwards.'*

Where there were duplicates of sounds (due to the original complexity of the mixes), the designer would often send *'one [object] to more important effects and the rest to less important.'* This was particularly pronounced for the 'Protest' scene mix, Figure 10.4, where there are numerous layers of crowd as well a different versions of background ambience. For example, of the three crowd foreground tracks (left, middle and right), only the middle track was assigned to MEDIUM IMPORTANCE with the remaining left and right versions set to LOW IMPORTANCE.

The main challenge encountered in assigning the NI metadata for the 'Protest' scene was setting the ambient crowd to be loud enough to convey the crowd's enormity whilst still maintaining intelligibility of the bank employees' speech. Considering the mix from the perspective of a hearing impaired individual, the designer remarked that a hard of hearing listener *'would probably struggle to hear where there's so much crowd going on on top of the speech [...] but she wouldn't get a sense of how scary it could be.'*

It was noted that many of the challenges encountered in the process did not come from the concept of assigning the importance itself but were due to the complexity of the session (having been originally mixed for surround sound with height). This meant that many objects had duplicates that would not normally occur in a standard radio drama: *'I mean a normal drama that I'd do isn't object-based, but it would be quite easy to apply these rules to it I think.'*

The narrative importance concept

The sound designer noted that narrative importance was not something specifically thought about when mixing, in her own words, *'Generally I sit at the end of [a mix] and say, "Let's make it as detailed a mix for the listener as possible so the listeners on really good*

headphones have a really good listen"'. However it was evident that making the NI metatdata assignments became quicker for the sound designer throughout the duration of the workshop and that the underlying concept was not incongruent with her mixing process.

This was particularly evident in the naming of the group functions in the plug-in over the course of 'The Turning Forest' mix. Initially, the groups were descriptive of the physical qualities of the sounds and types of objects (e.g. 'water' and 'ice') but as the workshop progressed, the groups were increasingly defined by the narrative role of the objects (e.g. 'FX, non-essential effects' and 'music, important').

In 'The Turning Forest', the sound designer only set one piece of music to HIGH IMPORTANCE, with all the remaining music being assigned to MEDIUM IMPORTANCE. In 'The Turning Forest', when the boy meets the creature, the creature smiles and his teeth turn into *'a wonderful string instrument'*. It was assumed by the author and other researchers involved that the piece of music associated with this plot point would be the most important piece of music. However, the sound designer felt that this music was only MEDIUM IMPORTANCE, as the Narrator describes the monster's musical smile. The most important piece for the sound designer was the 'spooky strings': *'that bit of music, that's more important than the other bits because it says something is about to happen.'* In particular, as there was nothing else to convey that part of the story, the designer considered this more important.

She also highlighted that the importance of the different objects changed based on the audience and intended use of the content. The 'Protest' scene was originally commissioned as a piece to demonstrate the creative capabilities of surround sound with height, making the sound of the helicopter flying over important as *'it was trying to show off what could be done with height'*. However, when mixing the stereo mix for a general audience, *'the helicopter is probably less important'*, as it doesn't play a narrative role.

The designer noted that initial expectations were that it would be more difficult to make the NI metadata assignments for 'The Turning Forest' than it would be for the 'Protest' scene but after the workshop noted that it was the opposite. The 'Protest' scene was challenging to assign NI metadata for due to *'how important the effects were in this one for telling the story, the background crowd and stuff'*. On the contrary, for 'The Turning Forest', she expected the sound effects to be *'really important, you've got the monster, but actually you can dial all that down because you get the story from the narration.'*

The sound designer felt that four was definitely the correct number of levels of narrative importance, as she noted *'I think that's a good amount because you need the one (*HIGH IMPORTANCE*) to peek through because they're telling part of the story or they're a signpost.* If there were fewer than four categories she felt that the lowest importance category would not be able to be sufficiently attenuated to give the intended benefit for hard of hearing

listeners. The designer reported being happy with the way that the end-user control linearly transitioned between the categories, and the level of attenuation which was applied to each category.

The designer also mentioned that a potential advantage of the narrative importance approach would be removing the need to compromise between mono mixes and the full desired mix for someone listening on headphones: *'Sometimes we are having to bring effects down, or music, to make sure the speech is crisp on the mono mix and actually I'd like it a bit louder'*. She felt that by giving the end-user greater control, the producer would have greater creative freedom.

Integration into production workflows

The sound designer felt that the workflow to author the metadata would likely not add to production time. If the content was recorded and tracks laid with this application in mind; *'I would have grouped my effects by* ESSENTIAL *in the same way these are, stuff that can go to three, stuff that can go to two and that is not extra work it's just a bit more careful planning.'*

In fact, she felt that the narrative importance approach would not only have the potential to deliver more creative freedom but make aspects of her current workflow easier. Specifically, she felt it was a useful tool for going from a surround or immersive mix down to stereo, stating that she felt: *'the end product of this has ended up better than the stereo bounce down'*, as it provided a *'better balance'* through the process of *'making the narrative more important.'*

She noted the challenges the current NI metadata assignment plug-in presented. The first was its inability to deal with stereo objects, requiring NI metadata assignment for a left and a right version of each object. The second was the inability to view the groups and the individual objects in the same window. It was indicated that a plug-in which resolved both these problems would have significantly streamlined the workflow.

10.2.3 Discussion

The results of this case study allow us to begin to answer the question posed in the introduction to this chapter.

Is the concept of narrative importance commensurate with production staff's own prioritisation strategies for audio objects?

Whilst the sound designer indicated that she did not directly consider the audio objects in her mix based on their narrative importance, from the workshop it was evident that the

narrative importance approach could be integrated into the way she developed her mixes. She adapted quickly to the concept and mentioned on multiple occasions that many parts of the NI metadata assignment were akin to processes she already used. In particular, she noted that checking at the IMMERSIVE and NARRATIVE ends of the scale were very similar to the checks she already does on her mixes to ensure they translate well to mono or poor quality speakers.

It was also noted that the narrative importance approach simply formalised some of the processes she already undertook in the mix. For example, an established technique she, and other sound designers use in radio drama, is to introduce an audio element to set the scene then reduce its level once the scene is set to give dynamic range to other audio objects. The narrative importance process allowed her to assign higher importance to the signposting audio elements, formalising this established technique.

It was suggested at the end of Chapter 9 that ESSENTIAL be used only for speech to give end-users greater ability to turn up the dialogue compared with other elements. The sound designer was not told this, but selected to only assign dialogue or human vocalisations to the ESSENTIAL category. This suggests that this restriction on category use may make sense and be acceptable to sound designers.

Is it feasible to integrate NI metadata acquisition into current production workflows?

The sound designer who participated in this case study indicated that she felt it would be feasible to integrate NI metadata acquisition into her workflow. She felt that the number of levels of narrative importance was appropriate and that it could streamline some aspects of her existing workflow and improve her creative freedom. She developed an iterative workflow involving object metadata assignment for all objects, audition of the mix, and then adjustment. However, it was evident that for this to be viable and not add to production time, two things would be required:

- Improved metadata acquisition tools, to allow easy auditioning throughout the assignment process.
- Consideration of the narrative importance application from the beginning of the production.

Another suggestion from the results of Study One in Chapter 9 was that, when speech was not present, the control could behave differently or revert to the original balance of elements. If changes in the gain of HIGH IMPORTANCE and other lower objects, based on the presence of speech, was implemented, this would make splitting up objects like the effort noises of Nathan unnecessary and further simplify the production workflow.

10.3 Study Three: Production staff survey

Study Two demonstrated that the concept of narrative importance is commensurate with an individual sound designer's conceptualisation of audio objects. Furthermore, it has showed that integration of NI metadata assignment would be feasible in her workflow.

Having established this for a single sound designer, Study Three now aims to determine whether these results generalise to other production staff. Furthermore, if this concept does generalise, it aims to determine whether production staff assign NI metadata in the same way or whether it varies based on professional experience.

To enable this study to gather data from the broadest range of production staff possible, a methodology based on qualitative survey items and an online NI metadata assignment task were implemented. The following section describes the design of the survey items and the NI metadata assignment task. Results of the survey items and task are then summarised, concluding with a discussion of the results with reference to the key research questions posited in this chapter.

10.3.1 Methodology

As demonstrated previously in this work (Chapter 3), online survey methods are effective for gathering data from large and diverse participant pools. Utilising a survey method, two types of data were collected: demographic and professional experience data, and closed- and open-ended questions about narrative importance and workflows.

To fully explore how different production staff would assign NI metadata, an additional task was included in the study's methodology. This task required participants to assign the importance of audio objects for an excerpt from 'The Turning Forest' using an online interface.

Participants were first directed to the survey which was hosted on onlinesurvey.co.uk. At the beginning of the survey, an introduction to the task, the project, and the technology was given. Once consent had been attained and a memorable ID for the participants generated, participants were asked to answer the demographic and professional experience questions. Following this, they were directed to a page hosted on a BBC R&D server where they would complete the categorisation task. They were prompted for their memorable ID, to allow for the data to be linked with their survey responses. They began the categorisation task by listening to 'The Turning Forest' radio drama in full and they had to reach the end of the drama before they could advance to the task. After completing the task, they were directed back to survey where they were asked closed- and open-ended questions about narrative importance, workflows, and their experience of completing the task.

The following sections describe the data collection for each part of Study Three, beginning with the survey items and followed by the NI metadata task.

Demographic and professional experience

Ten closed-ended questions were used to gather key personal and professional facts about the participants. The full list of demographic and professional experience questions can be seen in Table 10.3. Demographic data collected included their age, the country in which they spend most of their time working and whether they identify as hard of hearing. Five closed-ended questions then addressed the medium and genre in which they primarily worked, the type of organisation they worked for, the production role they most commonly undertook and the number of years experience they had in audio production. These questions were designed to characterise an individual's professional experience to allow for its potential effect on the results from other survey items and the NI metadata assignment task to be evaluated. Particularly, the question about the type of organisation for which they worked (large broadcaster, production house or freelancer) was designed to determine whether the established routines of a large broadcaster affected the results.

Finally, two further questions asked about their experience with objet-based audio (OBA) productions and whether they had heard 'The Turning Forest' radio drama before. These questions were designed to explore whether familiarity with the format and the content had any effect on their survey and task results.

NI metadata and workflows

Participants were asked seven questions about their conceptualisation of narrative importance, how they thought it might integrate into production workflows and their experience of undertaking the NI metadata assignment task. These can be seen in Table 10.2. Two of these questions address the narrative importance concept generally and are noted as C1 and C2.

Two questions explored participants' thoughts about how easily NI metadata acquisition might work in their own production workflows and whether they thought the number of narrative importance levels was appropriate. For the former, they were asked to rate how difficult they thought a hypothetical metadata acquisition workflow would be and then suggest any other ideas they might have for workflow integration. These questions are noted as WF1 and WF2.

Finally, three questions investigated how the participants found the difficulty of the task and what criteria they used to make their decisions. Motivated by the results of the focus group in Chapter 9, a question was also included about whether respondents felt the presence

of visuals would influence their metadata assignment; these questions are noted as T1, T2 and T3.

To complement the explanatory power of the demographic and professional experience data collected in the previous section, open-ended free text questions were added after each question to allow participants to explain their response in more detail. Participants were also given a final free text box in which they were given the opportunity to leave any other feedback.

NI metadata task

The aim of this task was to determine whether different production staff assign NI metadata in the same or different ways. To do this, participants were asked to assign the metadata for an excerpt of 'The Turning Forest', allowing comparison between the production staff who participated against the assignments of the original producer gathered in Study Two.

Optimally, this task would be conducted in a similarly ecologically valid environment as used for Study Two. However, to reach a broad range of producers required conducting the experiment online, consequently compromising ease of access with ecological validity.

Participants were asked to wear headphones whilst completing the task and indicate what make and model of headphones they were using. Headphone listening was selected, even though many participants were likely to have monitoring speakers available, as it minimised the potential for the room acoustics and ambient noise to affect participant decision making.

This section first outlines the excerpt of 'The Turning Forest' selected for the task, followed by the design of the online task interface.

Content The excerpt selected from 'The Turning Forest' was the 100 sec segment which begins with the protagonist Nathan chasing his friend Amelia into the forest before losing her and meeting the monster for the first time. There are 23 audio objects in this segment.

This excerpt was selected as it was inclusive of some of the key objects which the original sound designer discussed in Study Two. These included:

- The 'Spooky Strings' object which the designer considered the most important piece of music.
- The music accompanying the monster's smile.
- The sound effects indicating the creature's first appearance.
- The effort noises from the children which the original producer set to both ESSENTIAL and HIGH IMPORTANCE.

Table 10.1 Instructions for the NI metadata assignment task, which were shown on the task's online page above the interface shown in Figure 10.5.

Instructions
We are asking you to categorise sounds from a segment of the radio drama you just listened to based on the sound's importance to understanding the story. For each sound you should select one of the following categories using the radio buttons on the right: – Essential (red), – high importance (yellow), – medium importance (blue), – or low importance (green), based on how important you feel it is to following the content's narrative. Each object has a bar showing the periods when the object is active. You can navigate the track by clicking on these bars. Each object has a preview button below its name (left hand side) allowing you to listen to a representative, soloed clip of the sound (this will pause the overall mix). You can start and pause the track using the Play/Pause button.
Personalised Playback
There are three different playback options demonstrating a personalised audio implementation which would allow the user to adjust the complexity of the mix based on their hearing needs or listening scenario. The three playback options alter the sounds' reproduction levels based on the category you have assigned them to. The options are: – **High Complexity:** This is the original broadcast mix. All objects are reproduced without level alteration. – **Moderate Complexity:** The sound elements most vital (category essential) to understanding the narrative are increased in volume, the level of category high importance remain unchanged, whilst less important objects (category medium and low) are slightly attenuated. This mix is designed for listeners with mild hearing loss or listeners in a noisy environment. – **Low Complexity:** The most vital sound elements to the narrative are further increased in volume whilst low importance objects are heavily attenuated. This mix is designed for listeners with moderate to severe hearing loss or listeners in very noisy conditions. The playback at each setting of the control, and whether the mix sounds appropriate for the above scenarios, can be used to help inform your categorisation of the objects. The level adjustments applied to each category can be seen on the right of the level meters.

Table 10.2 Qualitative open- and closed-ended survey items used in Study Three, addressing key concepts relating to narrative importance, integration of NI metadata assignment into workflows and experience of completing the NI metadata assignment task.

ID	Questions about narrative importance concepts	Type of response
C1	Is the importance of a sound to the narrative something you consider when you mix?	Yes/No/Unsure
	– *Please elaborate on your answer.*	Free text
C2	Rate how you would feel about the audience being able to control the volume balance of objects in the mix based on their listening needs?	1 (Comfortable) → 5 (Uncomfortable)
	– *Please elaborate on your answer.*	Free text
ID	Questions about narrative importance workflow	Type of response
WF1	Do you think the number of categories was appropriate?	Yes/No/Unsure
	– *If you answered No, please explain why.*	Free text
	You work for a broadcaster and they have implemented the personalisation control demonstrated in the categorisation task allowing the end-user to switch between high, medium and low complexity mixes in their home. To allow for the necessary metadata to be collected, you have been asked to create four narrative importance buses in all the mixes you deliver and ensure all objects are routed to one of these buses.	
WF2	Rate how easily the above scenario would integrate into your current production workflow	1 (Easy) → 5 (Hard)
	– *Please explain why you gave this rating.*	Free text
	– *Can you think of an easier way of acquiring this metadata during the production process?*	Free text
ID	Questions about the NI metadata assignment task	Type of response
T1	Generally, how easy or hard did you find it to assign an importance category to each sound object?	1 (Easy) → 5 (Hard)
	– *Please explain why you gave this rating and any specific examples of objects you found particularly easy or particularly hard to categorise.*	Free text
T2	What criteria did you use to decide the importance level of each sound object?	Free text
T3	Would you change how you categorised the sounds if there was an accompanying video or a visual display?	Yes/No/Unsure
	– *If so, how would your categorisation change?*	Free text

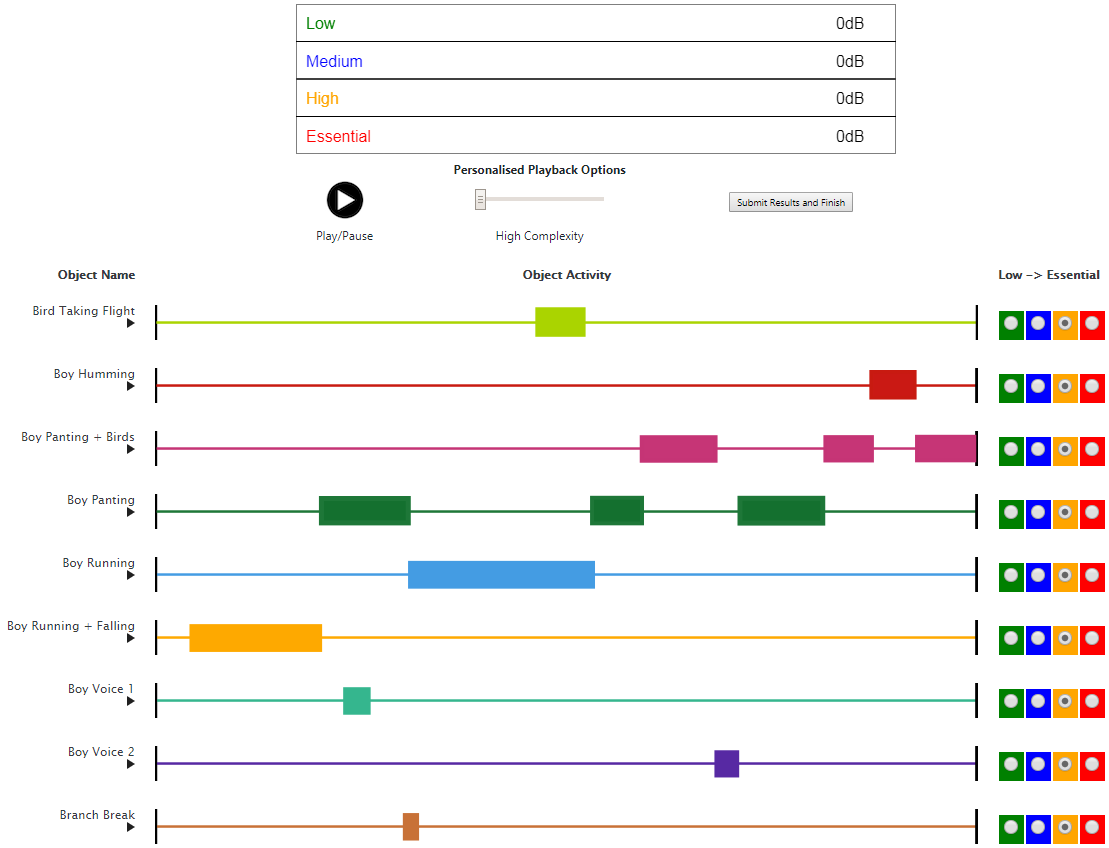


Figure 10.5 Interface for the online NI metadata assignment task, showing the 'Personalised Playback' control and the assignment interface for the first nine audio objects from 'The Turning Forest' excerpt, which includes a time line for each audio objects with the blocks indicating periods the object is active.

100 seconds was considered an appropriate length for the excerpt as it allowed participants to listen multiple times without the overall length of the task becoming too long.

The names assigned to some of the audio objects was altered from those used in Study Two as names such as 'Nathan Essential' would bias the participants. All the audio clips in 'Nathan Essential' and 'Nathan Non-essential' were renamed with descriptors of what they contained such as 'Boy Running' and 'Boy Panting'.

To ensure that they were familiar with how the objects would fit into the overall story line, participants were required to listen to 'The Turning Forest' in its entirety before beginning the task.

Interface The interface for conducting the NI metadata assignment was designed to be simple but with the key functionalities that production staff would expect to have in a DAW,

including the ability to skip to specific sections of the content and being able to solo objects.[3] The interface can be seen in Figure 10.5. The description used to introduce participants to the task and the functionality of the interface can be seen in Table 10.1. These instructions could be toggled off to allow the interface to occupy the entirety of the page.

Each object in the mix was listed alphabetically down the left-hand side of the page and down the right-hand side of the page there were four radio buttons for each object to allow the narrative importance level to be set. Each object would default to being assigned to the MEDIUM IMPORTANCE level, represented here in orange.

For each object, participants would be able to listen to a segment of the object solo-ed, using the play button under the name of the object. This allowed them to recognise the particular sound associated with the object name. A time-line was given next to each object's name and coloured differently for each object to help distinguish between them. To further aid identification of the object, the periods of the excerpt where the object was present were indicated by coloured blocks on the transport bar for each object. Each object had an interactive transport bar, allowing participants to skip through the content to a particular part they might want to audition.

To audition the mix as the end-user would be able to, a reduced version of the end-user prototype control was implemented. This can be seen in Figure 10.5 termed 'Personalised Playback Option'. Unlike the prototype in Chapter 9 and the interface used in Study Two, this control was not a continuous scale. Instead for ease of implementation, the control allowed playback at three levels, high, medium and low complexity, described in Table 10.1. These corresponded with an end-user input of 0. 0.5 and 1.0, as defined in the gain laws in Section 9.3.2, respectively. At each level of the control, the corresponding attenuations/gains applied to each category of objects was shown at the top of the page; the control would default to the 'High complexity' mix (seen Figure 10.5, with all categories showing attenuation of 0dB).

10.3.2 Results

Participants

Participants were recruited through professional organisations such as the Association of Motion Picture Sound (AMPS) and through internal communications in the BBC. Participants were also recruited through external promotion of the study via the University of Salford and BBC R&D's social media accounts.

[3]The interface for the online NI metadata assignment task was developed collaboratively by the author and Dr. Matthew Paradis. Dr Matthew Paradis completed the coding for the interface, as well as the data logging.

Thirty-four participants were recruited who identified as audio production or mixing professionals. One participant's results were omitted as their responses to the NI metadata assignment task were unchanged from the default settings and they indicated in their survey that they felt that the interface had not functioned correctly for them.

Demographics and professional experience

The responses to each of the closed-ended demographic questions can be seen in Table 10.3. Participants could select more than one answer in response to the first 3 questions.

It can be seen that the majority of participants worked in television production (42%) followed by radio (36%). Just over half of participants work as freelancers (53%), with most of the remainder working for a large or national broadcaster (36%). Documentary, drama, and music were the most common genres participants worked in and most identified their main role being that of a sound mixer (39%) or dubbing mixer (18%). There were 73% of respondents who were familiar with the concept of object-based audio (OBA), with 42% having worked on an OBA production before.

The majority of respondents were based in the UK (28), with the remaining from the United States (2), Australia (1), Germany (1) and Ireland (1). On average they had 21.8 years experience (median 21.0 years), with the experience normally distributed. This indicates that the results should be representative of a wide range of experience levels.

Participants had a mean and median age of 44 years. None of the participants identified as having hearing loss, though one participant opted not to respond to the hearing loss question.

The recruited participants represent a varied range of roles and experience within the production staff community. The high number of years of experience in the group helps give confidence that the respondents comments on workflow integration are likely informed by significant real-world experience.

NI metadata and workflows

This section discusses the results of each of the survey items sequentially, first addressing any closed-ended responses to the items and then summarising the key themes of the free text responses. The implications of these results for each topic is then discussed.

The narrative importance concept The first question in Table 10.2 directly addresses the first aim of this Chapter: *Is the concept of narrative importance commensurate with production staff's own prioritisation strategies for audio objects?* In response to this question,

Table 10.3 Closed-ended survey items addressing professional experience and personal demographics of respondents, with the number of participants selecting each response noted. Participants could select more than one option for the first three questions.

Questions	Responses	Counts
What medium do you most commonly work in?	• Television	14
	• Radio	12
	• Film	9
	• Other	6 – *Front of house (1), Theatre (1), Music (3), Student (1)*
What genre/s of content do you most commonly work on?	• Documentary	18
	• Drama	13
	• Music	12
	• Sports	5
	• Lifestyle	4
	• Comedy	4
	• News	2
	• Other	4 – *Commercials (2), Technology (1), Speech-based programmes (1)*
Which of the following best describes the organisation you currently work for?	• Freelancer	17
	• National/Large Broadcaster	12
	• Large Independent Production House	1
	• Other	4 – *Education (2), Theatre company (1), Research Institute (1), Post Production Facility (1)*
Which of the following best describes the majority of the work you do?	• Sound Mixer	13
	• Dubbing Mixer	6
	• Sound Recordist	4
	• Producer	2
	• Teacher/ Lecturer	1
	• Track Laying	1
	• Other	6 – *Radio Operations Sound Engineer (1), Sound Supervisor (1), Technology and systems design (1) 'Research + Mixing + Listening' (1), Student (1), 'most of the above' (1)*
Have you ever worked on an object-based audio production before?	• Yes	14
	• No, but I am familiar with the concept	10
	• No and I am unfamiliar with the concept	9
Had you heard 'The Turning Forest' prior to commencing this survey?	• Yes, once	3
	• Yes, more than once	14
	• No	16

100% of respondents said that the importance of a sound to the narrative was something they considered when mixing.

Respondents were asked this question after having already been introduced to the concept of narrative importance in the NI metadata assignment task. Whilst this may have biased the results, it also ensured that when the question was asked, all participants had an understanding of what was meant by the phrase *narrative importance*. Furthermore, this was such an overwhelming positive result that even accounting for some bias, it can be concluded that generally the narrative importance concept is comprehensible to production staff.

Furthermore, the free text responses from participants make it clear that this concept was reflective of thought processes and prioritisation which already exists in their workflows. One participant notes that if a sound is *'key to the story line I make sure its leveled [sic] to match its importance'*[4]. Another stated, *'you've got to prioritise on these things so that you can keep the soundscape extremely vibrant but not to the point of overloading the listener or drowning out the narrative'*.

Their comments also echo the many roles of sounds outlined by television soundscape theory in Section 9.2.1, with sounds taking on the roles of *'signposting and driving emotions'*, telling *'the audience how to feel about what they are seein [sic]'* and conveying parts of the narrative that *'can only be told through sound – events that are off-screen, for example'*.

End-user personalisation Participants were asked to *rate how comfortable they would feel about the audience being able to control the volume balance of objects in the mix based on their listening needs.* The results of this can be seen in Figure 10.6. It can be seen that the responses are heavily skewed towards end-user personalisation being acceptable to production staff with more than half giving the rating 1 – comfortable and no participants selected 5 – uncomfortable end of the scale.

To explore the respondents' rationale for rating assignment, a word cloud was generated from their free text responses which be seen in Figure 10.7. This was generated using the package `wordcloud` package in `R` [267]. The same approach was also taken in stemming and presenting the text data as in Chapter 3, with the text data first manually cleaned to remove typographical errors and to introduce consistency between abbreviations (for example, all instances of SFX and sound effects was resolved to SFX). Cleaning was also performed using the `tm` package in `R` [268] to remove punctuation and stop words and to stem the text to their base word stems; this resulted in 310 words with an average frequency of 1.9.

A Chi Squared Goodness of Fit (with a null hypothesis that the words were uniformally distributed) was undertaken ($[\chi^2 = 625.0, df = 309, p < 0.01]$). Calculating the maximum

[4]Levelled here refers to the process of structuring the gains of objects within an audio mix

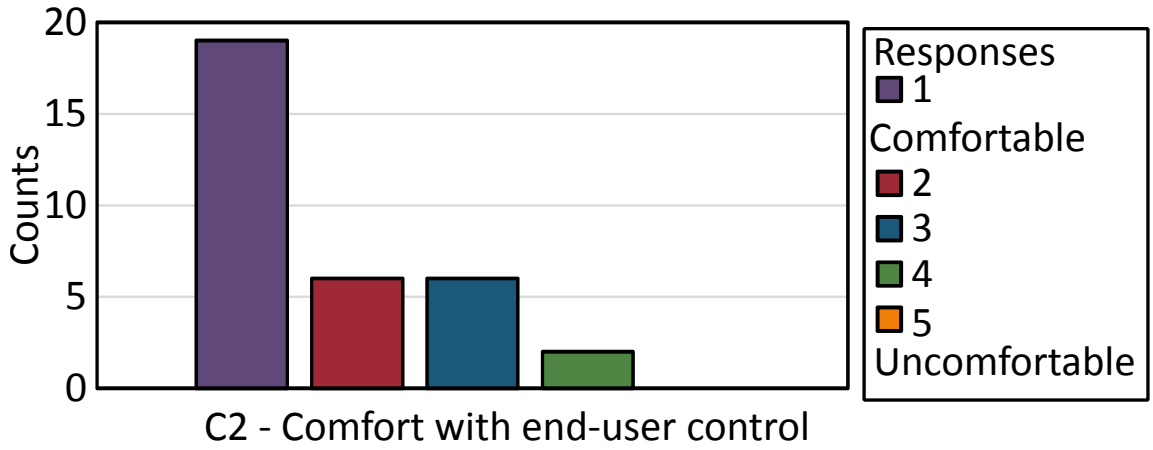


Figure 10.6 Responses to the question C2 from Table 10.2: *Rate how you would feel about the audience being able to control the volume balance of objects in the mix based on their listening needs?*, showing the majority of respondents are comfortable giving end-users control of audio objects.

level for significance at the level $\alpha \pm 0.05$, giving a significance level of $C_{sig} > 4.7$. The lower bound for significance was not calculated as it is only the key themes of interest in this analysis.

As in Chapter 3, larger font sizes indicated higher frequencies of word occurrence. Words depicted in colour (rather than black text) had frequencies of occurrence significantly higher than chance.

The audience was clearly at the heart of the production staff's opinion on personalisation, as seen in Figure 10.7. It is seen by a large proportion of production staff as a positive for audiences: *'if they need to tweak content to enjoy/get the most out of it then they should be able to'*, *'if it means that the content becomes accessible to a wider audience, this can only be a good thing'* and *'It is 2018. Let the viewer personalize his or her content!'*.

Many recognised that they already have little control over end-user reproduction: *'From [the] day the very first EQ was put in a radio, the audience have been able to alter our mix to their taste'* and that *'this is just an extension of accepting the reality of the end-user not necessarily hearing the mix as you intended.'*

It was evident that the majority of production staff surveyed are happy for users to have greater control over the content they consume, either through the narrative importance approach or other personalisation methods. However, these responses were not without their reservations, with some production staff concerned that *'The user might not know the best mix'* or that by changing the mix their *'audience would miss out on what I intended them to feel'*. Some felt though that these concerns were mitigated by the original mix still being available to the end-user and that *'as long as the story comes across and the audio engineer*

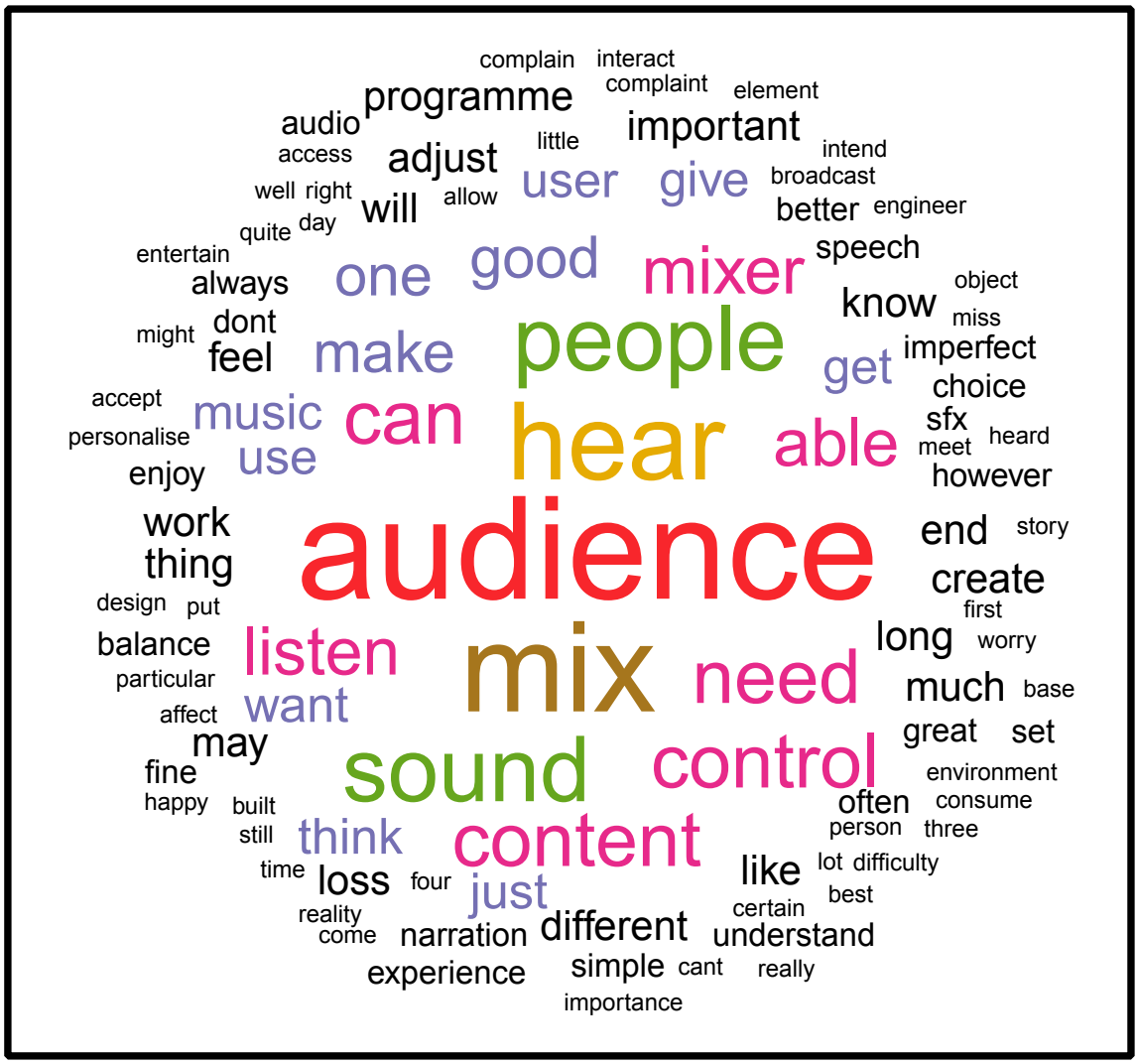


Figure 10.7 Word cloud of the most frequent words identified when participants were asked to elaborate on the question: *Rate how you would feel about the audience being able to control the volume balance of objects in the mix based on their listening needs?*. Words identified at a significantly higher frequency [p<0.01] are noted.

is the one making the call on whats [sic] important, this should work well'. Others expressed that they felt end-users would not be interested in engaging with the content in an interactive way as *'most people don't want to join in but are happy consuming'* and that *'they would listen as it comes and not bother fiddling around with the balance'*.

It is also seen by many as a positive for the production staff themselves: *'It's win-win as more people get to hear my work'* and that personalisation could be *'liberating for the content creator, who could make a mix more like the one they love without having to worry*

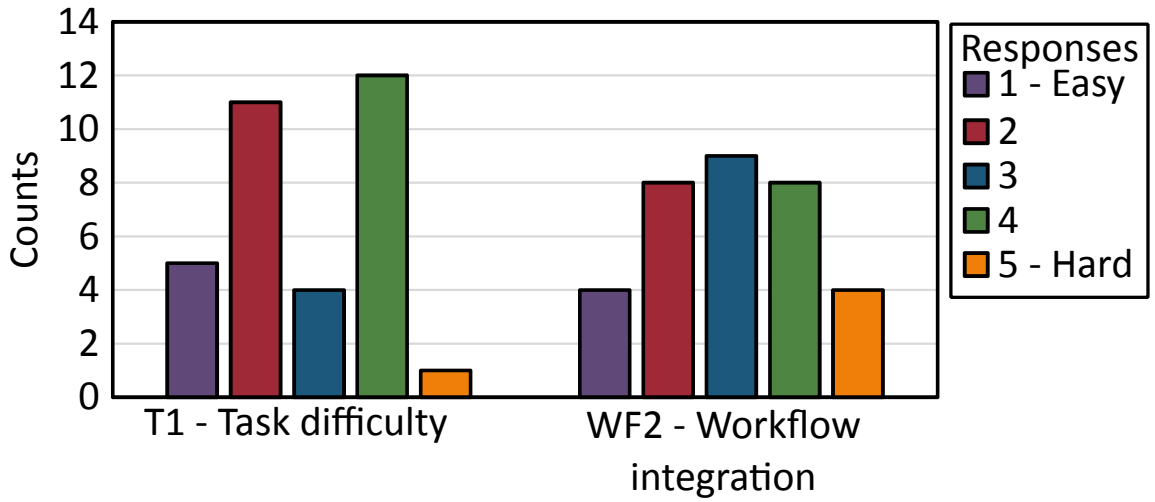


Figure 10.8 Responses to the question T1 from Table 10.2: *Generally, how easy or hard did you find it to assign an importance category to each sound object?* showing varied difficulty experience completing the task and responses to question WF2: *Rate how easily the above scenario would integrate into your current production workflow?* referring to the workflow described in Table 10.2, showing a normally distributed level of perceived difficulty.

quite so much about those with hearing difficulty or in a noisy environment'. This directly echoes the opinions expressed by the sound designer in Study Two.

Workflows The responses to WF2 and T1 can be seen in Figure 10.8. Both questions asked participants to rate their response on a five point Likert scale from 1 (Easy) to 5 (Hard). The distribution of responses for WF2 is normally distributed, with a mode and median of 3. This suggests that the proposed workflow is viable, at least in some production staff's current workflows whilst integration might prove more challenging for others.

A Pearson's Chi Squared test was undertaken to determine whether the demographic and experience factors of the participants affected their responses to this question. No significant factors were identified. That there was no significant influence from professional experience factors may indicate that the participant pool was not large enough to identify distinct trends. It may also suggest that there was greater variation between individuals' workflows than there was between the "standard" workflows of a particular type of content.

Participants were asked to elaborate on their rating of difficulty. Approximately a third of participants responded that integration would be easy, once they had acclimatised to it as the *'decision making would become easier over time'* and that initially it *'would take some time but once set up, not too hard'*. Many participants indicated that their workflow already involved grouping the mix into stems and this would just present a different way of grouping those stems. A further quarter of the participants indicated that whilst conceptually easy,

they felt it would introduce a significant time overhead into the post-production process and that it would adapt better to drama where post-production times are longer. Two participants indicated that they felt it would integrate easily enough as long as the necessary technical support was supplied (e.g. a way of *'quickly monitoring the final mixes during a live event'* or a *'dedicated VST'*).

Contrary to the participants who felt this would be a simple variation on their current workflow, some felt that the proposed approach would work directly against their current expected deliverables. One participant felt that the narrative importance mix would *'need to sit apart from the final mix'*. Participant identified other challenges that this workflow might cause were group processing, such as dynamic compression and reverb: *'I think reverb would be a problem because I usually apply it across the final mix. If somebody didn't hear a low-priority sound, they shouldn't hear the reverb from it either'*. How it might integrate into live production was also raised, as well as the challenge of the director having time to sign off on all the different versions of the mix.

Thiry-one participants responded to the secondary question *Can you think of an easier way of acquiring this metadata during the production process?* Of those, more than one third (12) responded, 'No', they could not think of an easier process. A number of alternative approaches were suggested including:

- Exporting content in the Audio Definition Model (ADM) format with metadata about bus routing or ability to tag.
- Utilising artificial intelligence to assign importance.
- Tag a clip directly.
- Importance automation in the DAW.
- Notes in the DAW.
- Use existing Dialogue and music and effects (M&E) buses and give end-users control over them.

The difficulty of integration of any NI metadata assignment process ranged from needing to be *'far simplier [sic]'* or *'anything that is one click or key stroke'* to the opinion that the proposed process *'could be implemented almost immediately, with no real time, or cost implications'*.

Two participants also brought up contrasting opinions about who should be authoring the metadata, with one participant noting *'I think the mixer is a good place to source the info, espeicaly [sic] as they have allready [sic] spoken to others (like the director) to obtain similar info for thier [sic] mix.'* In contrast, another participant suggests that importance

should be assigned earlier in the production process: *'having the importance level decided in the pre-production phase'*.

The free text responses indicate that, due to the large variety of workflows employed by production staff, the difficulty of integration also ranged from very simple to very difficult. Particular challenges exist in how to manage signal processing which currently occurs on grouped sounds in a mix, like compression, and how to manage the acquisition for particular types of content, like live content. Overall, it appears that with the right tools integrated into the DAW and with processes adaptable to different individuals and content types, acquisition of NI metadata could be integrated into much of the current production workflows.

Quantisation of scale Question WF2 from Table 10.2 explore whether the four level scale developed in Chapter 9 is considered appropriate by production staff. The majority of respondents (79%) indicated that yes, they thought the number of levels was appropriate. Of those remaining, 4 thought there should be fewer categories, 2 wanted more and 1 was unsure. Of those who wanted fewer categories, most suggested three would be sufficient: 'ESSENTIAL*, not, and middle of the road'*. One worked as a dubbing mixer for commercials, one as a Producer for Documentaries, one in Radio Operations Sound Engineer for speech programming and one was a technologist. It is possible that their roles influenced their opinion and that for commercials and speech-based radio programming, fewer categories would be more appropriate. Both respondents who indicated they would like more categories worked in music, with one respondent suggesting that the current categories represent too large a change in level and *'more categories, with smaller level changes would be preferable.'* For the vast majority of production staff surveyed, the proposed four level scheme was considered appropriate.

Difficulty of task Participants were asked how difficult they found the NI metadata assignment task. The responses are seen in Figure 10.8, with the median response being 3 and the mode being 4. Responses were distinctly bimodal, with the majority of participants finding it somewhat hard or somewhat easy. A statistical analysis using Pearson's Chi squared analysis was again undertaken and showed that the demographic and experience factors of the participants did not significantly affect participants' responses to this question.

Exploring the text responses indicated where some of the greatest challenges came from, with many participants indicating their lack of familiarity with the task or the content increased the difficulty of the task: *'Because of my lack of familiarity with the content it took as [sic] while to get a feel for what was important in the material'*. For this reason, they felt it was important that *'the person to categorise the content is the producer/mixer*

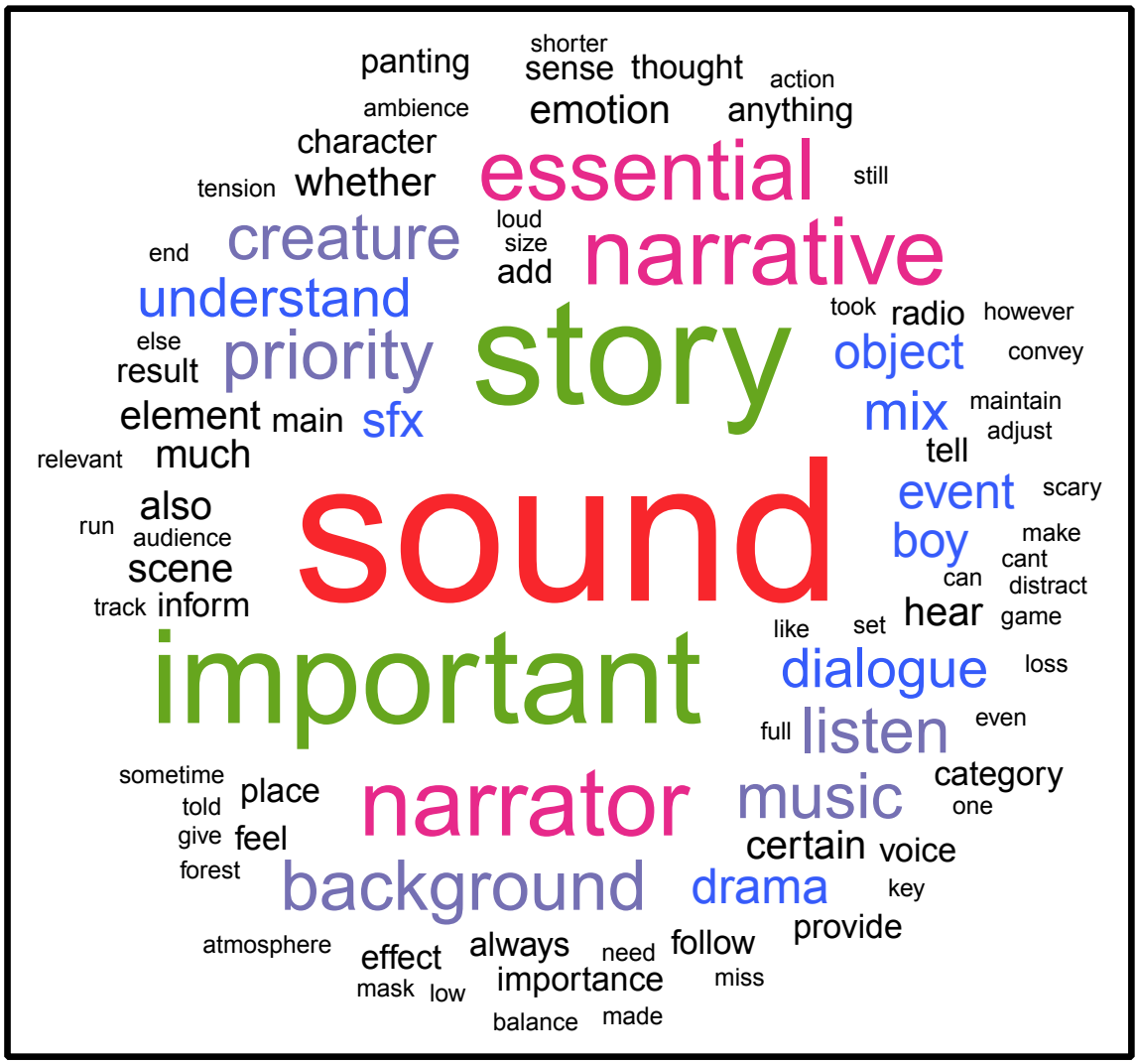


Figure 10.9 Word cloud of the most frequent words identified in responses to question T2 from Table 10.2: *What criteria did you use to decide the importance level of each sound object?* Words identified at a significantly higher frequency [p<0.01] are noted.

rather than a third party'. Many indicated that they felt once they had acclimatised, the task would become much easier: *'Once I had decided on my own rules to follow it was easy to implement'* and it would have been *'even easier if I had been involved in the production process'*. This again reflects that production staff feel that assignment of NI metadata should be part of the production process, not a post-hoc process; furthermore it echoes the sentiment that the process should be completed by production staff involved in the content creation. Some found it challenging because it involved balancing many more audio objects than they were experienced with: *'Concentrating on the mix of anything above 8 tracks is not*

something I normally do'. Like the original sound designer, one participant commented that *'it's difficult to assign low importance to any object'*.

Many participants indicated that objects which they felt to be at the extreme ends of the scale (ESSENTIAL and LOW IMPORTANCE) were easy to assign. Sounds which were in the middle two levels presented a much greater challenge, in particular *'the kids voices and the creature voices were harder to categorise'*. This may be in part due to familiarity with the content, with objects like ambience and narration occupying more standard roles in the content, whereas the sounds of the children being more specific to the intent of the original sound designer. This also suggests that completion of this process by individuals involved in the production would likely streamline it.

Criteria Participants were asked to explain in free text responses what criteria they used to make the NI metadata assignment decisions. The free text responses were first analysed utilising a wordcloud to facilitate identification of key themes, as for the opinions on personalisation. This can be seen in Figure 10.9. A Chi Squared Goodness of Fit test (with a null hypothesis that the words were uniformally distributed) was undertaken ($[\chi^2 = 872.8, df = 276, p < 0.01]$). Calculating the maximum level for significance at the level $\alpha \pm 0.05$ which gave a significance level of $C_{sig} > 4.7$. Words which were identified significantly more often are shown in coloured text. The lower bound for significance was not calculated as words identified significantly less often were not considered to have explanatory power.

Sound, importance, and story are the highest frequency words with many participants indicating that their criteria was simply *'revelance [sic] for understanding the story'* and whether *'the story could make sense without it [the sound]'*. Also key to many participants' criteria was ensuring that the narration was prominent and that there was *'clarity of dialogue, above all else'*.

The importance of conveying emotion, particularly via music, was highlighted by a number of participants. One participant initially set the music as low importance and found: *'the sense of trepidation disappeared'* and they had to reinstate it *'until i was happy that 'atmosphere'* was back. Others noted that some of the *'music beds were important to conveying the mood'*; however, *'not all of the short stings were as essential'*. This is again congruent with the underlying idea of narrative importance with not all objects that share physical qualities, such as music, having equal importance in conveying the story. As with the original designer's assignments, one participant noted the 'Spooky Strings' object, in particular, stating that they are crucial *'to building tension in the scene, when the narrator can't achieve that alone'*.

Effect of visuals Participants in the focus group in Chapter 9 noted that the inclusion of visuals would have an effect on which sounds they wanted included in the mix and which they felt would be less important. To explore this idea with the production staff cohort, the final task related question in Table 10.2 asked whether they thought their categorisations would change if the content had visuals. Participants were also split over whether visuals would have affected how they assigned the NI metadata. Just over half of respondents (19) indicated that visuals would affect their assignment of NI metadata. A further 9 were unsure and the final 5 participants said No.

There were two main themes present in the free text indicating how they thought visual would affect the NI metadata assignment: first by raising the importance of some LOW IMPORTANCE sounds and secondly, by lowering the importance of some HIGH IMPORTANCE sounds. For some, the sounds they assigned lower importance like a 'Branch breaking' and 'Bird taking flight' would need to be more prominent if accompanied by a corresponding visual to ensure audio-visual congruency. Conversely, the presence of a visual image to convey things such as the size and scariness of the monster would mean that sounds conveying its approach would not need to be as loud or imposing. One participant sums up this dichotomy by saying, *'Visual cues may be enough to get across the story, so the sound might not be necessary. On the other hand, if I see an action it must be accompanied by a SFX otherwise it would feel wrong'*.

NI metadata task

This section summarises and discusses the results of the NI metadata assignment task for the excerpt of 'The Turning Forest' excerpt. To initially evaluate the similarity between the different participants' assignments, Fleiss's Kappa for assessing the reliability of agreement was calculated. This measure is defined as:

$$\kappa = \frac{\overline{P} - \overline{P}_e}{1 - \overline{P}_e} \tag{10.1}$$

where the term $[1 - \overline{P}_e]$ *indicates the degree of agreement that is achievable above chance and* $[\overline{P} - \overline{P}_e]$ *gives the degree of agreement above chance which was achieved.*

Fleiss's Kappa is equal to 1 when total agreement is achieved and equal to zero where there is no agreement. For the NI metadata assignment, Fleiss's Kappa was calculated to be $[\kappa = 0.11]$ indicating a very low level of agreement between participants.

Whilst all participants stated that narrative importance was something they considered when they mixed (Section 10.3.2), it can quickly be determined from this initial evaluation of agreement that how they assign that importance differed immensely.

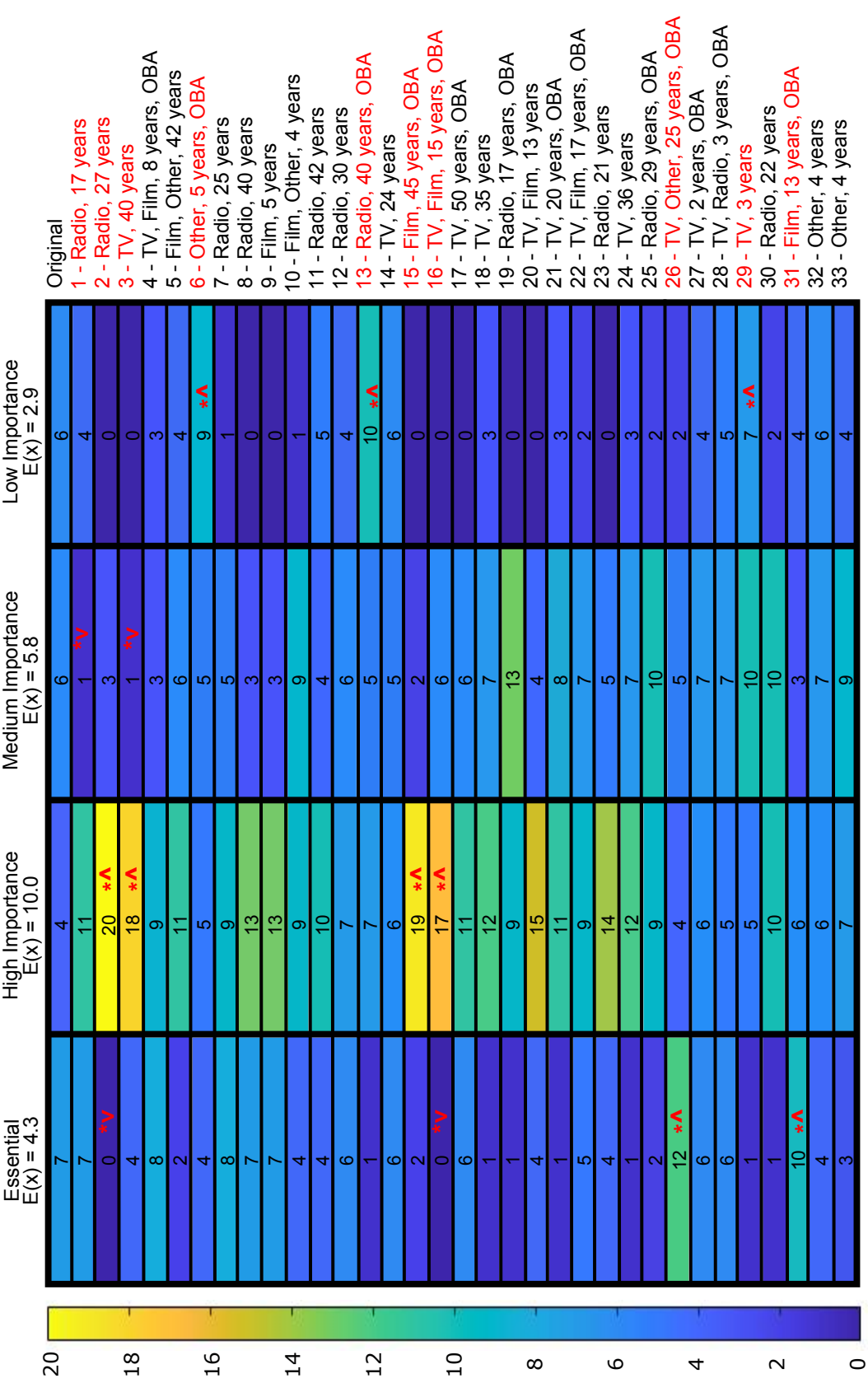


Figure 10.10 Heatmap showing how many objects each participant assigned to each level of narrative importance in the NI metadata assignment task. Expected number of objects for each category noted as E(x). Years experience and main medium worked in are noted beside participant ID number. Participants with significantly different proportions of objects in any level are noted in red text, with the significantly different level noted by a red *^ for significantly higher and *v for significantly lower proportions of assigned objects.

The data was further interrogated to determine where these variations occurred and whether the professional experience of participants explained some of this variance. A visualisation of how many objects participants assigned to each narrative importance level was generated and can be seen in Figure 10.10. The original sound designer's assignment was included for comparison and the medium and years of experience for each participant was noted for reference.

To determine which individual participants assigned objects in a significantly different manner than others, a Pearson's Chi Squared test was performed on the data (inclusive of the original producers assignments) giving $[\chi^2 = 255.8, df = 99, p < 0.001]$. More than 20% of the comparisons had an expected frequency less than 5 which indicated that the Chi Squared value may be over-estimated [270]. However, as the Chi Squared value was more than double the threshold required for significance and the minimum expected frequency for all combinations was > 1, it is likely that the effect of this over-estimation was only small. [5]

Inspection of Figure 10.10 indicates that the largest range of values exists in the HIGH IMPORTANCE level, ranging from 4 objects assigned to 20 (87% of the objects). For most participants, the HIGH IMPORTANCE level represents where the majority of objects were assigned (as indicated by the expected frequency of 10.0); this is quite different to the original sound designer's assignments which had only 4 objects in the HIGH IMPORTANCE level.

Ten participants were seen to assign the objects significantly differently (denoted by red text in Figure 10.10). Inspection of the standardised residuals indicated which combinations of participants and categories were significantly different and these are noted with a red *^ for combinations with significantly higher assignments and a *v for significantly lower combinations.

Two participants did not utilise the ESSENTIAL or LOW IMPORTANCE levels, only assigning objects to either HIGH IMPORTANCE or MEDIUM IMPORTANCE (Participant 2 and Participant 16). This resulted in these participants having significantly less assignments in ESSENTIAL and significantly more in HIGH IMPORTANCE. Participants 26 and 31 assigned 10 or more objects to ESSENTIAL, though they then roughly evenly distributed the remaining objects across the other categories.

Ten participants did not use the LOW IMPORTANCE category, reflecting the comment in Section 10.3.2: *'it's difficult to assign* LOW IMPORTANCE *to any object'*. This sentiment was also initially expressed by the original sound designer in Study Two. However, after auditioning the mix further, she finally assigned 6 objects to LOW IMPORTANCE. This

[5] It should be noted that whilst the original sound designer only assigned the 'Narrator' and 'Nathan Essential' into the ESSENTIAL category due to the decomposition of 'Nathan Essential' into short audio objects with neutral descriptive names, the number of objects assigned to ESSENTIAL for the original sound designer increased from 2 to 7.

suggests that a common initial response to NI metadata assignment was to avoid setting objects as LOW IMPORTANCE but through greater familiarity with the process and better tools to audition the mix, production staff become more willing to set objects as LOW IMPORTANCE.

To explore whether the professional experience of the participants explained any of the variance in the patterns of assignments, an ANOVA using the percentage of objects in each level was performed. For each narrative importance level, the Anderson Darling Test was used to determine whether the percentage of objects in that level were distributed normally. The LOW IMPORTANCE level was seen not to be normally distributed, so the ANOVA was only applied to the ESSENTIAL, HIGH IMPORTANCE and MEDIUM IMPORTANCE levels. Nine binary categorical variables describing individuals professional experience were used in the ANOVA which were:

- 'What medium do you most commonly work in?'
  - TV production experience
  - Radio production experience
  - Film production experience
- 'Which of the following best describes the organisation you currently work for?'
  - Employment as a freelancer
  - Employment as part of a large or national broadcaster
- 'Which of the following best describes the majority of the work you do?'
  - Mixing experience (either identifying as sound mixer or dubbing mixer)
- 'Have you ever worked on an object-based audio production before?'
  - Experience in object-based audio productions (response 'Yes')
  - Familiarity with object-based audio concepts (responses 'Yes' or 'No, but I am familiar with the concept')
- Familiarity with 'The Turning Forest'

As participants could give more than one response to the question, 'What medium do you most commonly work in?', responses to this question had to be separated into three separate binary variables. The responses to questions about the organisation also had to be separated into binary variables and only 'freelancer' and 'Large, National broadcaster' were retained as they were the only responses identified enough times for any potential effects to be observable. Similarly, the role of 'mixer' was the only role selected from the responses to 'Which of the following best describes the majority of the work you do?' and sound and

dubbing mixer was combined in order to have substantial numbers of participants in each category. Finally, experience with OBA was split into production experience and familiarity with the concept to determine whether each individually affected the patterns of assignment.

For the ESSENTIAL category, none of the factors were seen to be significant. For the HIGH IMPORTANCE and MEDIUM IMPORTANCE, the ANOVA results are summarised in Table 10.4 with the significant factors noted in bold.

Table 10.4 Analysis of variance (ANOVA) results between the proportion of total audio objects assigned to the HIGH IMPORTANCE and MEDIUM IMPORTANCE levels and the professional experience factors.

	HIGH IMPORTANCE			MEDIUM IMPORTANCE		
	Sum Sq	F	p			
TV	0.024	0.65	0.43	0.001	0.08	0.78
Radio	0.026	1.5	0.23	0.007	0.5	0.49
Film	0.202	5.44	**0.03***	0.054	3.6	0.07
Freelancer	0.001	0.01	0.981	0.008	0.5	0.49
Large Broadcaster	0.003	0.08	0.78	0.005	0.3	0.57
Mixer	0.064	1.72	0.21	0.038	2.5	0.13
OBA Production Experience	0.157	4.22	0.052	0.061	4	0.056
OBA Familiarity	0.027	0.73	0.4	0.07	4.6	**0.04***
Familiarity with Turning Forest	0.008	0.22	0.64	0.0002	0.02	0.9
Error	0.855			0.346		

It can be seen from Table 10.4 that in assigning HIGH IMPORTANCE, professional experience in film production had a significant effect on how the objects were assigned. Post-hoc inspection using Tukey's Honestly Significant Difference test showed that those with experience in film were more likely to have a larger percentage of the audio objects in the HIGH IMPORTANCE than those without film experience. Experience in OBA production was also borderline significant [p=0.052], with post-hoc analysis showing that those who had never worked on OBA productions were also more likely to have a larger percentage of their objects in the HIGH IMPORTANCE.

For MEDIUM IMPORTANCE, familiarity with the concept of OBA had a significant effect on the assignments. Post-hoc analysis showed that those who had familiarity with the OBA concept would assign fewer objects to MEDIUM IMPORTANCE. Again, OBA production

experience was borderline significant [p=0.056], with those who had worked on an OBA production assigning more objects to MEDIUM IMPORTANCE.

These results, coupled with Figure 10.10, suggest that those with experience in film tended to assign more objects to high importance rather than spreading out the assignment across the categories. Given that the audio mixes in film are often louder and more complex, as a controlled cinema reproduction environment is assumed, this may explain why film production staff were more likely to assign more objects to higher importance in the mix.

Conversely, knowledge of OBA concepts and experience in OBA productions resulted in a more even spread of objects across the different levels. This suggests that better understanding of OBA's purpose and features, and experience of using it, made participants more likely to exploit the range of values and functionality of the technology.

A final analysis was undertaken to determine how consistently individual audio objects were assigned to particular levels of importance. This is visualised in Figure 10.11. Each audio object is represented in the narrative importance level where it was assigned by the majority of participants (modal level). The percentage of participants who assigned it to that category is also noted.

It can be seen from 10.11, that the only object which had over 67% agreement on its importance was the Narrator. For all other audio objects, there was significant variation in where the audio objects were assigned. This indicates that not only did the percentage of objects assigned by participants to each level of importance differ, but the exact objects they assigned there varied as well.

Comparing the narrative importance level most commonly assigned to each object with the level assigned by the original sound designer showed that only 8 of the 23 objects were assigned to the same level (shown in Figure 10.11 in red text). Half of these objects are at the end points of the scale, with all the objects most commonly assigned by respondents into ESSENTIAL and LOW IMPORTANCE being assigned to those same categories as the producer. Notably, the 'Creature Smile' music which the original sound designer decided was less important than 'Spooky Strings', was here most commonly assigned as a higher importance. This echoes the earlier comment that objects at the extremes of the scale are simpler to assign than those in the middle. It appears there is also more agreement at the extremes than the middle. It re-affirms further that how individuals assign NI metadata is very individual.

10.4 Discussion of Studies Two and Three

There were two key questions which these studies aimed to address which were:

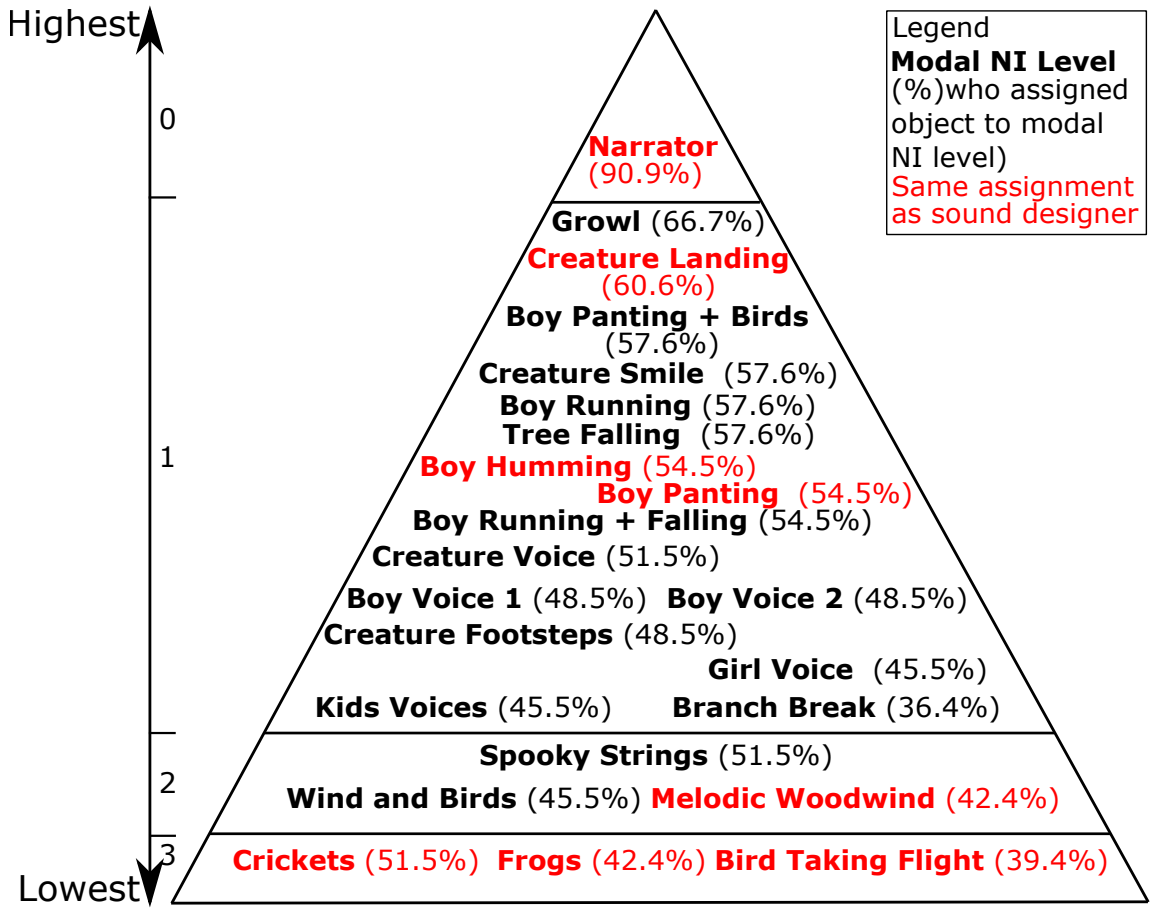


Figure 10.11 Hierarchical visualisation of the modal narrative importance level assigned to each audio object in 'The Turning Forest'. The percentage of respondents assigning each object to the modal level noted in parentheses and red text denotes objects which were assigned most commonly by participants to the same narrative importance level as assigned by the original sound designer.

- Is the concept of narrative importance commensurate with production staff's own prioritisation strategies for audio objects?
  - Does this conceptualisation differ between production staff?
- Is it feasible to integrate NI metadata acquisition into current production workflows?

It is evident from both studies that the concept of narrative importance is something which makes sense to production staff. Study Two highlighted that the concept is already present in the way, at least some, sound designers mix. This is further corroborated by Study Three where 100% of respondents said that the importance of a sound to the narrative was something they considered when they mixed. The parallels drawn in Study Two between the sound designer's existing processes and the narrative importance process further support

this idea. The results of these studies indicate that the concept of narrative importance works within, not against, production staffs' existing ideas of audio object priorities.

Whilst the concept is understood by varied production staff, the results of Study Three show that the way the concept is implemented is very individual. Both studies suggest some key factors which influence this including familiarity with the content, the medium in which production staff have experience (particularly film) and familiarity with OBA processes. However, the factors covered here are likely to be only a small portion of those which influenced participants' opinions on how audio shapes the narrative. It appears, in the words of Cohen, that it is not only the viewer that *'is actively engaged in constructing a narrative'* [355]. The content creator is also constructing the narrative based on their personal experiences and environment.

For the majority of production staff, it seems that four levels of narrative importance are an appropriate number. The supremacy of dialogue appears to be the one point of agreement between most production staff. The results of Study Two suggested that the approach put forth in Chapter 9, that the top category be used only for speech to maximise the speech to background ratio achievable, would be acceptable. However the large variety of objects which some participants included in the ESSENTIAL category here suggest that this may not be acceptable to all producers.

The studies here also show that integration of NI metadata assignment into production workflows is feasible. For the sound designer in Study Two, it would be a simple integration and indicated that any extra effort required would be balanced by the other aspects of the mixing process that it streamlined. Many recognised that the process integrating a new workflow would be a challenge and take additional time, but once a new workflow was established it would become easier and quicker.

The proposed workflow in Table 10.2 was recognised by many as viable, with one participant proposing that it *'could be implemented almost immediately, with no real time, or cost implications'*. Suggestions for alternate workflows focused primarily on factors that would make the process as simple as possible. Some participants highlighted that this might be a suitable application for artificial intelligence. [6] From the varied assignment results here however, it appears that any successful AI approach would need to be semi-automated with the human production staff in the loop. In particular, it seems that assignment of objects at the extremes of the scale, where there is greater agreement, could be automated whilst the more difficult decisions in the middle of the scale would always require human input. As

[6] An initial feasibility study on the application of machine learning based classification for automatic assignment of NI metadata has been undertaken in collaboration with BBC R&D, Queen Mary University London, and the author. It is described in [388].

stated by one participant, the system would work well as long as the audio engineer was *'the one making the call on whats important'*.

Many found the task of assigning NI metadata challenging due to their *'lack of familiarity with the content'* and that it would have been easier if they *'had been involved in the production process'*. This sentiment was also present in Study Two, with the sound designer indicating that the process would have been streamlined had the narrative importance concept been in mind from the beginning of the production. These ideas reaffirm the thesis put forth by Youngblood that *'Accessibility cannot be an afterthought. It needs to be part of the creative process.'*[105]

Other participants indicated that workflow integration would be more challenging, and in some cases work contrary to their existing deliverables. Other challenges were highlighted that have broader implications for OBA delivery, such as the application of group processing and the assignment of metadata in live workflows. These challenges are bigger than the narrative importance concept and should form the focus of future OBA research.

There is often the assumption that production staff will be resistant to personalisation and do not want to cede any control of the content to the end-user. Some participants referred to this idea in their free text responses suggesting their *'industry can get too precious about "the art"* and that *'sound engineers don't like giving the audience control, they think it will reveal imperfections'*. However, the opinions expressed here by production staff about personalisation largely indicate the opposite (though with the caveat that as participants opted into the study, anyone with strong opinions against personalisation was likely not to choose to participate). Figure 10.7 shows audience needs are core to the considerations of production staff. Furthermore, many identify greater ability to personalise content with an opportunity for greater creative freedom – being able to deliver a mix *'more like the one they love without having to worry'* that it will not be accessible to audiences.

Finally, as concluded in Chapter 9, determining the feasibility of this system for television content requires a study utilising real broadcast workflows and ecologically valid audiovisual content.

10.5 Chapter Summary

This chapter undertook two qualitative studies to determine whether the concept of narrative importance was commensurate with production staff's conceptualisation of audio object importance. Study Two focused on exploring these ideas in-depth with one sound designer whilst Study Three investigated how these conceptualisations differed across a cohort of production staff. The secondary aim of these studies was to determine whether acquisition

of NI metadata was feasible in current production workflows and, if so, what supporting technology would be required to achieve this.

The key conclusions drawn from these studies are:

- The concept of narrative importance works within, not against, existing ideas of audio objects importance held by production staff.
  - However, the way in which importance is assigned differs greatly between production staff and is influenced by their professional experience.
- Integration of NI metadata assignment into production workflows is feasible, but dependent on:
  - Being conducted by production staff,
  - Being conducted as part of production, not after it,
  - Having effective and simple tools available to facilitate it.

The results from the production staff surveyed here also show that assumptions that production staff are resistant to personalisation are likely unfounded. Many see it as liberating, giving greater creative freedom, and they feel that if personalisation allows them to reach a greater audience then *'it can only be a good thing'*.

Chapter 11

Large-scale evaluation of the narrative importance approach to accessible audio

11.1 Introduction

The first three studies in Part III of this work have indicated that an accessible audio system based on the narrative importance concept is a viable personalisation solution for end-users and a feasible approach for production staff. However, all studies concluded that proper evaluation of this approach required: ecologically valid audio-visual content, production within real production workflows and schedules, and evaluation by a large target cohort. This chapter implements such a study to conclude this work's investigation of the question:

Can a system be designed to allow end-users to control the balance between audio objects for dramatic content which is simple to use and preserves comprehension?

This chapter reports on a large-scale public trial of the narrative importance concept, production workflow, and interface called *'The Casualty Accessible and Enhanced (A&E) Audio Project'*. The study used a full episode of BBC Studio's 'Casualty' programme [389]. The content used for the study is first outlined, followed by a description of the data collected from participants. The production of the narrative importance mix is then discussed in detail, along with an analysis of the characteristics of the narrative importance mix.[1] A large public cohort participated in the study (n=6228). Responses about the function and value of the control were also gathered from a subset of the public cohort (ratings: n=353, feedback survey: n=273). These responses are summarised and their results discussed.

[1]The terminology adopted in this work for referring to narrative importance is to use the full term to refer to the concept of narrative importance or a narrative importance approach. The abbreviation NI is used for technical implementations – such as NI metadata assignment, NI prototype and NI control.

11.2 Study Four: Methodology

The three aims of this study are:

1. To explore the production of narrative importance content in real production workflows.
2. To evaluate how individuals interact with the end-user interface for narrative importance content.
3. To determine what value the narrative importance approach to accessible audio has for a broad range of normal and hard of hearing audiences.

In order to address the first of these aims, real broadcast content was selected and production of the narrative importance mix was completed with the programme's production team as a part of their mixing process for the episode. A description of the content is given in Section 11.2.1 and the production process is described in Section 11.3.

At the time of writing, object-based techniques were not widely used in broadcast systems or production workflows. To trial the object-based narrative importance approach with ecologically valid broadcast content, a hybrid approach, compatible with current production workflows, was required. Fundamental to this approach was developing a way of embedding the NI metadata into channel-based audio assets. The approach developed is described in Section 11.2.2 as is the implementation of the end-user interface to playback these assets.

To evaluate aims 2 and 3 of the trial, data was collected from members of the public from two sources:

- the BBC Taster site: eight demographic and feedback questions plus a rating of the technology on a five-point Likert scale.
- the standard media player: interaction data describing user behaviour with the control during playback (including the level at which the NI control was set and the level at which the volume control was set).

The methodology of this data collection is described in the final part of this section.

11.2.1 Content

In Chapter 3, respondents identified 'drama' as the most problematic genre for speech understanding. For this reason, the content selected for the study was a drama programme. The results of Chapters 9 and 10 further indicated that this should be a television drama to provide the study with ecologically valid results.

The enthusiasm and availability of programme production teams to be involved in the study influenced the selection of the content. This, along with the suitability of the content, led to the selection of an episode of BBC Studio's 'Casualty' for the study.

The 'Casualty' programme follows staff in the Accident and Emergency (A&E) department of a hospital in the fictional location Holby City. As it is a hospital drama series, set in a busy emergency ward, the show presents many dense auditory scenes. 'Casualty' is also the longest running emergency medical drama in the world, with a large existing viewership. It was envisaged that this following would aid in driving traffic to the study. 'Casualty' broadcasts weekly on BBC One's terrestrial transmission, a.k.a 'BBC One' and is also distributed via BBC iPlayer. Both these platforms provide stereo audio, closed captions (subtitles) and audio description.

Plot Synopsis

The specific episode selected was Episode 38 from Season 33 which has a run time of 49:25mins. This episode was primarily selected based on its focus on hearing impairment in the plot and showcasing of Deaf actors. It was felt that this would aid publicity of the study and engagement of the target audience with the content. The plot is summarised below to aid in interpreting analysis of the different scenes in the latter portion of this chapter.

Episode 38 begins with an attempted theft and motorcycle crash which results in two patients being admitted to Holby City A&E, one of whom is named Jason and is a Deaf character played by a Deaf guest actor. One of the main plot lines follows a nurse, Jade, and her interaction with Jason, diagnosing his condition and learning about his involvement in the accident. Jade is played by Gabriella Leon who is a hearing impaired cast member of 'Casualty' (and can be seen in 11.1). Also involved in the crash is innocent bystander Barbara who is also taken to the hospital and the episode follows her treatment.

Another major plot line follows character Danny, in the wake of her mother's death, pretending to be a paramedic and trying to treat Barbara. Ruby, who has previously encouraged her in this endeavour, is disciplined by her superior Jan. After Danny tries to get back in contact with Ruby, she has a car crash as she drives to the hospital to see her.

Dylan, a doctor, and Jade also take care of a patient (Magdalena) by applying larval therapy. These scenes are peppered by many sarcastic and dismissive conversations between Dylan and Magdalena about her role as a spiritual healer. Finally, she reveals that her day job is in employment tribunals focusing on inappropriate workplace behaviour highlighting the improper way Dylan has acted towards her.

Minor continuing plot lines are:

- The management of the emergency department by Ciaran who senior management have brought in to improve the department.
- Jacob's ongoing interaction with a woman who is claiming to be caring for his mother.

- Jason's brother Chris and his ongoing attempt to get Jason into a gang he is in.
- Jade's ongoing battle to be confident in her nursing role.
- Iain's return to work and adapting back to daily life after a suicide attempt.

Short scene descriptions, noting the location of the scene and the main speakers in the scene, can be seen in Tables 11.1 and 11.2. The episode is made up of 53 short scenes which intertwine the different plot lines. These are on average 55 seconds long, with the longest being 198 seconds (involving a key plot point where Danny is coaxed from her car at the site of a crash) and the shortest being a discussion between two doctors lasting 8.5 seconds.

11.2.2 End-user interface

To embed the NI metadata into the channel-based audio assets, the routing capability in a standard DAW was exploited to create stereo NI busses. This adopted a similar workflow to that described and evaluated by participants in Chapter 10. Each level was assigned its own stereo bus which could be manipulated based on the narrative importance gain laws and rendered to an eight channel AAC audio file. This narrative importance mix could then be played back using the end-user interface seen in Figure 11.1.

The interface has two parts, the user interface and the system with which the eight channel audio stream is decoded.[2] The user interface was developed as a plug-in to the BBC's Standard Media Player [390]. This media player is the underlying interface used across the BBC's platforms for reproduction of audio-visual media and supports addition of new user interface components using plug-ins. Development used the Standard Media Player for three reasons: it is ecologically valid, it maintains BBC brand visual identity and it makes the technology compatible with the BBC's online streaming service, iPlayer [391] On the study's webpage, the Standard Media Player was hosted within an iFrame which the user could initiate by clicking a button labelled 'Try it'.

Given the layout of the Standard Media Player, which has a horizontal bar for its media controls, a slider rather than a dial was selected. This ensured that the NI control was easily identifiable to users and could be controlled in the same manner as the other media controls on the Standard Media Player provides. The slider ranged continuously from 0.01 – 1 and names were given to the two extremes of the slider, as well as the central location, to aid user interpretation of the control. These names are different from those used in previous chapters to ensure that the language describing the control was simple and accessible. These were:

[2]The integration of the NI slider into the Standard Media Player was completed by Matthew Paradis, with the author's guidance and support from Robin Moore.

Table 11.1 Scene by scene summary of Episode 38, Season 33 with scene duration, whether the scene is set inside or outside, location and a short description on the key plot points and characters – Part I, scenes 1 - 26.

Scene #	Duration (sec)	Inside/ Outside	Location	Description
1	44	Multiple	Multiple	"Previously on 'Casualty'"
2	46	Outside	Street:busy	Attempted theft, moped crash
3	36.5	N/A	N/A	Intro Credits
4	57	Outside	Street:busy	Aftermath of moped crash
5	32	Inside	Office	Connie and Ciaran talk
6	31.5	Outside	Car/Street: quiet	Jacob sleeps in car outside his mum's house
7	57.5	Inside	A&E	Ciaran and staff discuss new initiative
8	34.5	Outside	Nondescript	Chris meets with gang members
9	44.5	Inside	Resus	Jason is brought into the hospital
10	41.5	Inside	Resus	Barbara brought in; Barbara, Archie and others talk
11	23	Inside	In House	Danny at home, after trying to treat Barbara
12	71	Inside	A&E	Dylan begins to diagnose Magdalena
13	85.5	Inside	Staffroom	Jade, Dylan and others talk about a nursing bursary
14	53.5	Inside	Resus	Jade, Jason and others talk incl. BSL
15	54.5	Inside	A&E	Dylan continues to diagnose Magdelena
16	30	Inside	Staffroom	Jan, Ruby and Iain talk about Ruby's involvement with Danny
17	36	Inside	House	Danny sits remembering her mother
18	54.5	Inside	Resus	Barbara, Archie and others discuss Barbara's state
19	37.5	Inside	Resus	In Resus: Jade, Jason and others treating Jason
20	13	Outside	Street:quiet	Chris followed by Police Car
21	46.5	Inside	A&E	Jade and Dylan discuss Jason
22	81.5	Inside	Office	Jan, Ruby and Ian talk
23	48.5	Inside	A&E	Dylan continues to diagnose Magdelena
24	39.5	Inside	Resus	Barbara, Archie and others talk
25	47	Inside	Resus	Jason moved back to resus after tests
26	58.5	Inside	House	Danny calls Ruby

Table 11.2 Scene by scene summary of Episode 38, Season 33 with scene duration, whether the scene is set inside or outside, location and a short description on the key plot points and characters – Part II, scenes 27 - 53.

Scene #	Duration (sec)	Inside/ Outside	Location	Description
27	21	Inside	A&E	Jade and Ethan talk about Jason
28	135.5	Inside	Resus	Archie tries to save Barbara
29	50.5	Outside	Street	Paramedics arrive at Danny's car crash
30	41.5	Inside	Resus	Barbara dies
31	24	Outside	Street: busy	Danny refuses to leave car
32	38	Inside	Hallway	Jan, Ruby and Iain talk, decide to attends Danny's car crash
33	115	Inside	A&E	Dylan continues to diagnose Magdelena
34	49	Inside	Resus	Dylan and Jade talk
35	40.5	Outside	Hospital	Chris talks to gang member
36	27.5	Inside	A&E	Dylan, Jade and Magdelena talk
37	198	Outside	Street: busy	Ruby arrives at Danny's car crash and coaxes her out of the car
38	97	Inside	Resus	Dylan and Jade talk about Jason's condition
39	17	Outside	Street: busy	Danny is loaded into ambulance at the crash scene
40	102.5	Inside	A&E	Ruby and others talk, Danny arrives
41	8.5	Inside	Resus	Dylan, Jade, Ethan and others discuss Jason's meningitis
42	41	Inside	A&E	Jan, Ruby and Iain talk about repercussions for Ruby
43	38.5	Inside	Resus	Jade and Jason's brother talks, Jason has a seizure
44	59.5	Inside	Resus	Archie and Ciaran talk about new initative
45	77	Inside	A&E	Magdelena suggests some improvements to Dylan's bedside manner
46	93.5	Inside	A&E	Ruby and Danny talk about how Ruby cannot see Danny anymore
47	33	Inside	Resus	Jason, Jade and brother talk
48	125.5	Inside	Staffroom	Ruby and Iain talk about Iain's mental health and Ruby's history with Danny
49	56.5	Inside	A&E	Ciaran, Archie and others talk about the data lost from the new initiative
50	62.5	Inside	Pub: busy	Paramedic Support group
51	103	Inside	Office	Dylan and Jade talk
52	64	Inside	Pub: busy	Paramedic Support group
53	37	N/A	N/A	Closing Credits

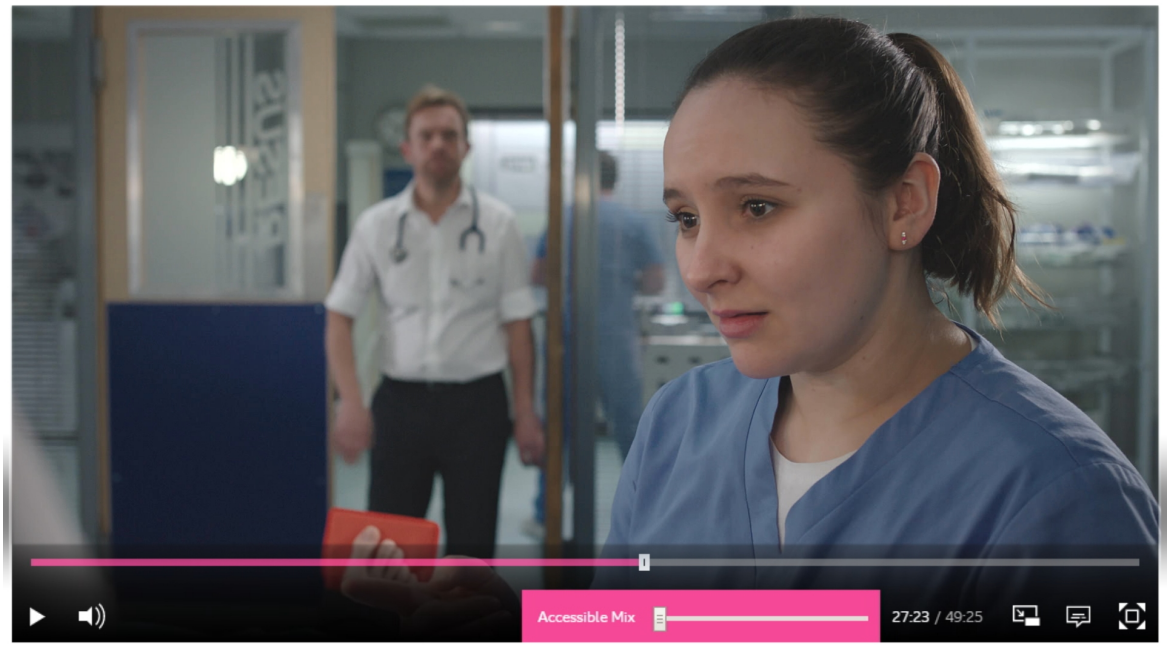


Figure 11.1 End-user for interface Study Four implemented as part of the BBC's Standard Media Player, with a screenshot from Episode 38 from Season 33 showing character Jade and with NI control slider highlighted in pink.

- TV MIX at the right hand extreme, corresponding with low end of the scale, previously termed the IMMERSIVE end
- ENHANCED, in the middle, corresponding with halfway
- ACCESSIBLE, at the left hand extreme, corresponding with top end of the scale, previously termed the NARRATIVE end

The slider defaulted to the TV MIX position. It was highlighted in pink when it was hovered over or interacted with (as with other UI components of the Standard Media Player).

To decode the eight channel AAC audio stream, the Web Audio API [392] was used (as for the public experiment in Chapter 10). This then applied the narrative importance gain laws depending on the position of the slider.

User interaction data

User interactions with the media player and the NI control were logged and were instantiated when the user first interacted with the iFrame [3]. However, the system only began logging

[3]The routine to log this data was developed collaboratively with Dr Matthew Paradis and the author, with the script to log and store the data scripted and implemented by the latter. All code to parse the data and all analysis of the resulting data was completed by the author.

data once they had reached the main episode content (either by watching or skipping through the pre-roll video).

Once participants had reached the main content, a log entry was stored into a buffer whenever they changed the slider value (and released the mouse). If the participant grabbed the slider and moved it around before releasing the mouse, only the final value when they released would be logged. The buffer could store up to three values which would be sent to the server and stored as .json files. When participants closed the iFrame, any unsaved data which remained in the buffer was also sent to the server. Whilst preferably a greater granularity of data would be collected, this approach was selected to ensure that the server was not overloaded in periods of high activity.

Four pieces of data were logged, in addition to a random session ID, which are described in Table 11.3.

Table 11.3 Summary of the different types of interaction data logged by the end-user interface, including the formatting of the data and its range of values.

Data type	Format/Range
Date/Time stamp	in the form 'YYYY-MM-DD T HH:MM:SS.SSS Z'
NI slider	unitless / [0.01 – 1] in increments of 0.01
Volume	unitless / [0 – 1] in increments of 0.1
Time in the program	seconds / [0 – 2965.099] in increments of 0.001

An understanding of how individuals interacted with the technology could be determined from the collected data. However, there were a number of systematic limitations to this data due to its method of collection which were:

- No initial time stamp was recorded when the log file was instantiated, preventing the calculation of accurate total interaction time.
- Up to two log entries could be lost if participants quit the tab or the browser rather than closing the iFrame. This also limits the accuracy of total interaction time calculations.
- Logs may not be unique and may represent an individual interacting with the interface on different occasions.
- Only final slider values were saved, meaning dynamic slider behaviour was only partially characterised by the data.

How these limitations were managed is described in Section 11.4.2. In addition to this, there was a problem with the data logged on the first three days of the pilot. A fix to this was put in place on the 12th June 2019 so only data from then on-wards is analysed in this work.

Feedback data from the Taster platform

Feedback data was collected to address the third aim: *to determine what value the narrative importance approach to accessible audio has for normal and hard of hearing audiences*. The platform used to distribute the study online was the BBC's Taster Platform which is designed to allow the public to trial prototypes of new broadcast technology and new types of broadcast content. This platform includes functionality to collect feedback data from consenting participants. This platform has a mandatory option for participants to give the technology a rating out of five stars and two default demographic questions (age and gender). In addition to this, the study's authors can add six multiple-choice feedback questions. These questions are limited to 40 characters for the questions (including spaces) and the multiple-choice responses are limited to 20 characters.

The questions which participants were asked can be seen in Table 11.6 in the results section (Section 11.4.1). Three demographic questions were added based on which factors were perceived as likely to affect respondents' experience of the technology: level of hearing loss, location whilst watching and whether they were regular viewers of 'Casualty'. Two questions about participants' opinion of the control were asked, as well as the default question about whether the BBC should 'do more stuff like this?' These questions asked whether participants noticed a difference based on the control and what that difference was. The specific wording of the questions and the response options can be seen in Table 11.6.

Participants who opted to answer these questions could answer as many of the total questions as they wished and also had the ability to skip questions. Their data was saved if they responded to at least one question.

11.3 Production of the narrative importance mix

The results in Chapter 10 suggested that the process of NI metadata acquisition should be part of the post-production process, rather than a post-hoc task. However, for this trial, it was requested by the post-production team that the narrative importance mix was produced after the broadcast mix was completed. This section begins by outlining the workflow adopted to create the narrative importance mix followed by a description of the production process for the mix.

After the mix was completed, peak limiting was applied to the stems to ensure they would be reproduced through the end-user interface at an appropriate level. The remainder of this section analyses the characteristics of the mix. This includes the distribution of signal power across the different NI stems and the effect of the NI control on the overall glimpse

Figure 11.2 Image of the Dubbing Theatre at BBC Studio's Roath Lock Studios in Cardiff, with superimposed screenshot from the Episode 38 from Season 33, showing character Jade [image courtesy of 'Casualty'].

proportion (GP) and hence, energetic masking, of the mix. A description of the GP can be found in Section 4.4.

11.3.1 Workflow methodology

The development of the workflow and production of the narrative importance mix was completed over a period of 13 hours. It was completed at the BBC Roath Lock Studios in Cardiff where 'Casualty' is regularly mixed (seen in Figure 11.2). The mix was primarily done by the 'Casualty' Dubbing Mixer, assisted by the author and with editorial input from the Producer and Post Production Supervisor.

Whilst a plug-in for auditioning narrative importance mixes in standard DAWs had been developed [393], the Pro Tools[394] implementation of the plug-in was not completed in time to be used in the mix. For this reason an alternate approach to auditioning the mix had to be developed based on the 'Casualty' Dubbing Mixer's existing workflow.

The narrative importance mix was achieved by creating four stereo NI busses within the existing Pro Tools session for the episode. These were routed through an AMS Neve DFC Gemini digital mixing desk and auditioned on PMC Monitors. Different levels of the

end-user control were auditioned manually using the corresponding channel strip for each NI bus.

The Pro Tools session template used for mixing this episode grouped the track into seven categories, which can be seen in Table 11.4.

Table 11.4 Summary of the audio groups used in the 'Casualty' Pro Tools template and description of their contents.

Group	Description
Dials/Sync	audio recorded on set at time of filming, primarily containing dialogue but also containing other on set sounds
Automated Dialogue Replacement (ADR)	any dialogue which was not captured suitably on set and required re-recording in post production
Music	score
Backgrounds/Atmos	any long term background sounds and recorded ambiences
Stereo FX	any spot effects which were added in post production, usually gained from a sound effect library
Mono FX	as above, for any mono spot effects
Foley	specially recorded sound effects

Each of these existing groups was stratified into a suitable number of tracks for each narrative importance level. This signal flow can be seen in Figure 11.3. Each group was addressed sequentially and all the NI metadata assignments for each group was completed over the whole programme before moving onto the next group. The majority of this process was auditioned with the gains corresponding with the ACCESSIBLE end (i.e. ESSENTIAL with +3dB gain, and MEDIUM IMPORTANCE with -12dB and LOW IMPORTANCE with -48dB attenuation respectively). A description of the types of sounds that were assigned to each narrative importance level and the rationale for this assignment are described and discussed sequentially in the following sections.

Essential

The NI metadata assignment process began with the Dials/Sync which was first edited to separate any other components which were not dialogue and did not overlap any dialogue. These were moved to a separate track in order to allow them to be treated as Foley or Spot Effects. The Sync tracks (with dialogue only) and ADR were then routed to the ESSENTIAL bus.

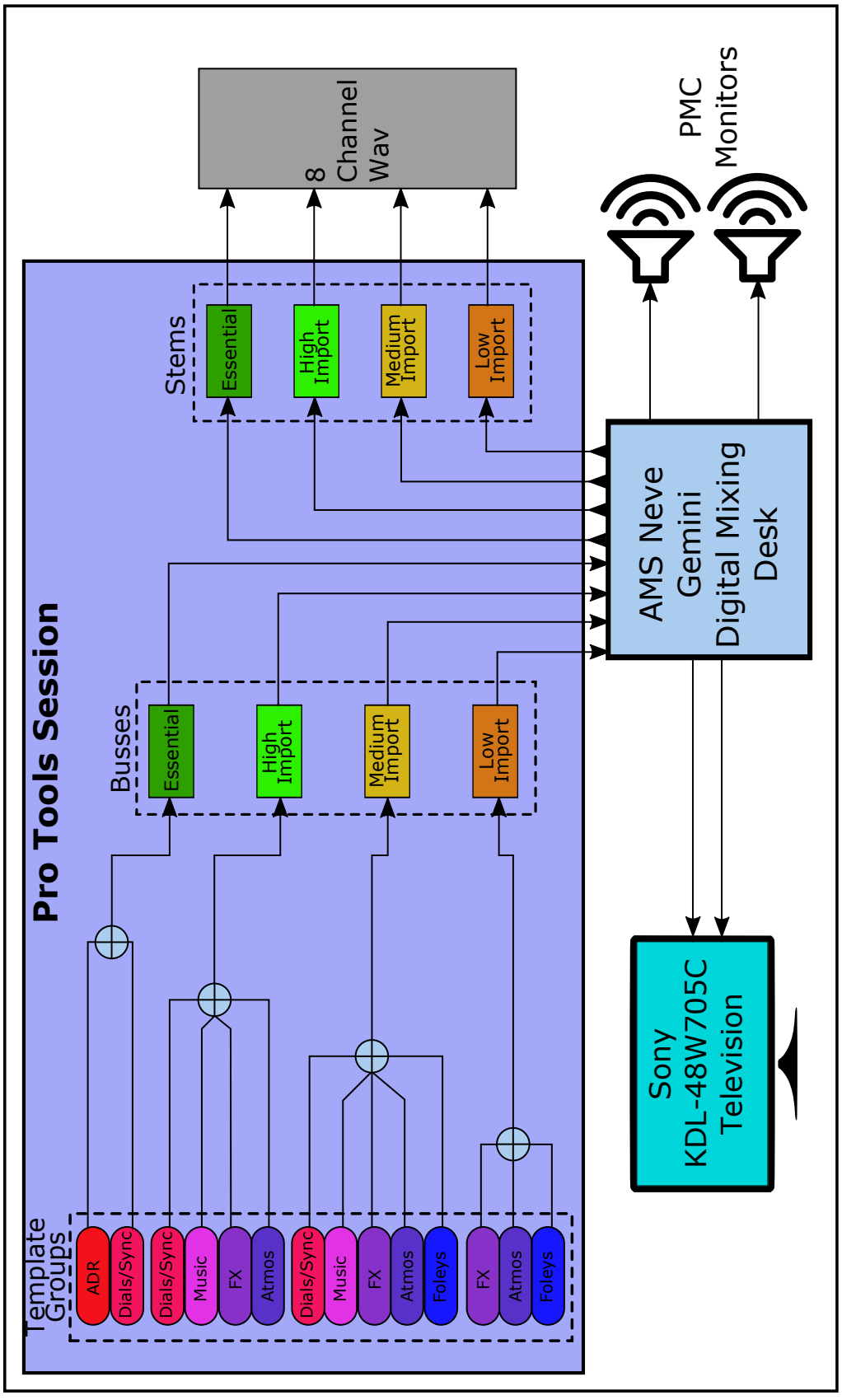


Figure 11.3 Visualisation of the signal flow for the developed NI metadata assignment workflow in the Dubbing Theatre at BBC Studio's Roath Lock Studios. For each level of narrative importance the original template groups (seen in Table 11.4 from which the objects for each level were sourced are noted, as well as the monitoring and rendering outputs.

Based on the outcomes of the initial user-experience testing in Chapter 9, only dialogue and sounds inextricable from the sync dialogue were routed to the ESSENTIAL level. Dialogue was inclusive of effort noises, like the gasping of an injured patient, and vocalisations which complemented the actors conversing in BSL.

Some characteristics of the sync sound, such as clothing rustle, became more apparent at the ACCESSIBLE end of the mix. It was noted that should the narrative importance approach become a regular workflow, greater attention may be needed in the recording process to these noises to ensure they were not captured in the dialogue.

High Importance

The sounds routed to the HIGH IMPORTANCE bus can be seen in Table 11.5. Approximately half of the scenes (25) had some sounds routed to the HIGH IMPORTANCE bus. The main types of sounds which were routed to the high narrative importance levels were music and important spot effects. The two main rationales for routing clips to HIGH IMPORTANCE were to draw attention to the sound or to ensure the emotion of the sound/scene was conveyed. For example, the squelching sound of the maggots (used in larval therapy) when Magdalena tries to walk and falls over was kept at HIGH IMPORTANCE (scene #36). This ensured that the disgust and awkwardness of the situation was conveyed. Some ambiences from the sync track were also routed to the HIGH IMPORTANCE bus as, due to the edits made to remove the non-speech sounds from the sync track, if these were attenuated too heavily, the edits would become obvious.

Similarly, most of the sounds in Scene 2 were kept at HIGH IMPORTANCE. This scene establishes the context for the majority of the episode's plot lines and has minimal dialogue and consists mostly of sound effects and suspense building background music. Given the importance of conveying the emotion and severity of the crash to the audience, the music, motorcycle sounds and many other street sounds were routed to HIGH IMPORTANCE. Subsequently there was minimal difference between the ENHANCED and ACCESSIBLE mix for this scene.

Only one piece of dialogue was not routed to the ESSENTIAL bus. This was a line spoken by the character Danny who was sitting in her car after a car crash refusing to leave (scene #31). In this scene, Danny's voice is intended to be muffled as the shot is viewed from the perspective of the paramedic who is outside the car. Given the creative intent, and the fact the line of dialogue is short, this was routed to the HIGH IMPORTANCE bus to ensure it remained muffled even at the ACCESSIBLE end of the slider.

There were four pieces of music which were used in nine separate instances across the episode. All except one piece of music were routed to the HIGH IMPORTANCE bus. The title

Table 11.5 Audio objects assigned to the HIGH IMPORTANCE bus by the dubbing mixer, noted by scene.

Scene	Audio objects
2	Motorcycle, phone, music, sound of the phone being stolen, moped crash sound effects, gasping
3	Title music
4	Gasping, ambulance arriving, sound of closing knife
6	Knocking (on car window), car starting
11	Danny putting down stethoscope
17	Music
20	Police car driving
26	Phone (calling Ruby), Ruby closing locker (where stethoscope was stolen from), music
27	Nursing rep walking, Jade movement's, replacing bandage
28	BSL vocalisations, Jason whimpering in pain, Barbara's heart monitor (indicating cardiac arrest)
29	Phone, stethoscope (being placed onto car seat), music, crash sound effects
30	Barbara's heart monitor
36	Sound effect of maggot being squashed, footsteps
37	Ambulance arriving, Danny's dialogue
38	Jade's footsteps (after deciding against orders to see Jason), Jason's vocalisation, Jason's whimpering, Door (bursting open)
41	Dylan's footsteps
42	Jan's footsteps
43	Jason having a seizure
46	Music,
48	Music, sound effects of sitting and standing, stethoscope (sound of placing on bench)
49	Computer error tone
50	Pub ambience (primarily from sync sound)
52	Music
53	Title Music

music is used in the intro and final credits. This music was the only audio object in the mix during the credits. It was set to HIGH IMPORTANCE such that it remain unchanged in the reproduction at the end-user.

The remaining pieces of music routed to HIGH IMPORTANCE were a bass-y, driving piece and an emotive, instrumental piece made up primarily of strings. The first of these is played behind the motorcycle crash in the second scene to build and enhance the tension of the scene.

The emotive string music was used in six instances across the episode in emotive scenes relating to Ruby, Danny, and Iain. In relation to Danny, the music is played under Danny crying in scene #17, Danny trying to contact Ruby in scene #26, Danny ending up in a car crash in scene #29 and in scene #46 where Ruby tells Danny they can no longer be friends. It is also played under the conversation between Iain and Ruby about his mental health and her interactions with Danny in scene #48, and in the final scene before the credits where Iain has decided to join Ruby at the paramedic support group. All of these instances of music were set to HIGH IMPORTANCE.

In all of the scenes with the emotive string music, the music plays an important role carrying the emotion of the scenes and linking the elements of this plot line together. In all except two of these scenes, #46 and #48, there was minimal, to no dialogue. In the scenes with dialogue, there is only speech, music and background room tone as the main audio objects. Given this sparseness, even for the scenes with dialogue the emotive string music was set to HIGH IMPORTANCE.

Only two audio clips had their absolute level altered in the mix. One of these was a section of sync dialogue which had birds in the background and the birds were reduced between the speech to ensure that when the dialogue was boosted by 3dB at the ACCESSIBLE end of the scale the birds were not unreasonably loud.

The majority of decisions about the narrative importance of sounds were made easily. It was evident however that a knowledge of the over-arching series plot lines were important to this decision making. An example of this was the sounds the movements of Ruby's stethoscope made (which Danny stole in a previous episode) being routed to HIGH IMPORTANCE. Keeping these sounds high in the mix drew attention to them and the plot points they represented.

Medium and Low Importance

Whilst working through the backgrounds/atmos group, there were many points at which the outdoor or hospital ambience consisted of multiple, layered, and similar audio clips. Where this occurred, often one clip would be left at the MEDIUM IMPORTANCE level and

the remainder routed to the LOW IMPORTANCE. This allowed some sense of location to be maintained whilst ensuring that non-essential background sounds could be heavily attenuated at the ACCESSIBLE end of the mix.

One piece of music was routed to MEDIUM IMPORTANCE. This is the music used in the background of the opening scene – *"Previously on Casualty..."* – which gives a catch up of previous events on the show and only has dialogue and background music. In this scene the background music was set to MEDIUM IMPORTANCE, so that it could be reduced but never entirely removed (as it is the only other audio element apart from speech).

11.3.2 Finalising the Mix

Once this process was complete, editorial input on the NI metadata assignment was gained from the Producer and Post Production Supervisor. The mix was then auditioned as it rendered in real-time. It was auditioned at the *TV Mix* end of the scale on a commercial television set (Sony KDL-48W705C) set with rear facing speakers. During this process a number of tracks which had been routed incorrectly, as a result of the group processing not being migrated correctly to the track level, were identified. Every time an error was caught, the rendering had to be re-started. As a result, rendering the final stems took three times longer than expected.

Ordinarily the final step of the mixing process before exporting the audio would have a -3dB limiter applied and the mix would be adjusted to meet the ITU BS.1770-4 standard for integrated programme loudness [94]. Given the variable reproduction levels of the four NI stems, this could not be carried out in the same manner as usual. To establish the range of the loudness values, integrated loudness was measured for both the combined stems with no gains or attenuations applied (TV MIX), and with the ACCESSIBLE mix gains and attenuations. A 2.6dB difference was found between the TV MIX and ACCESSIBLE mixes. Given that the loudness measurements are dominated by speech frequencies and at the ACCESSIBLE end the speech is increased by 3dB, the magnitude of this difference was close to what was expected.

The delivery specifications for BBC broadcast content insist upon the final mix loudness being -23 $\pm$ 0.5dB^{LKFS} [97]. Given the use of the content in for trial purposes, meeting these delivery specifications was not required. To try and bring the mix as close to the broadcast levels as possible, all stems were attenuated by 3dB and then passed through a ITU-R B.S. 1170-4 [94] compliant NUGEN Audio ISL True Peak Limiter [395] set to -3dB true peak. This gave an integrated loudness of -24.3dB^{LKFS}at the Full Mix and -21.7dB^{LKFS} at the ACCESSIBLE mix. This was then exported as a eight channel AAC file for playback in the Standard Media Player.

Workflow challenges

As identified in Chapter 10, this workflow presented a particular challenge for group processing. The clips in the Pro Tools session for 'Casualty', as in most post production content, are grouped based on type. This allows for group processing in the form of equalisation and compression can be applied to all objects in the group given their similar spectro-temporal characteristics. Whilst processing for equalisation and noise reduction in dialogue were (relatively) easily migrated from the group level to the track level, compression could not be. This was particularly challenging for the high and MEDIUM IMPORTANCE busses which contained music (for which the compression was most relevant). A compromise arrangement was found whereby a version of the type of compression usually used on music was placed on the HIGH IMPORTANCE bus and normal effects bus compression was place on the MEDIUM IMPORTANCE and LOW IMPORTANCE busses (despite the MEDIUM IMPORTANCE bus containing music).

11.3.3 Analysis of the Mix

This section investigates how the NI metadata assignment process and reproduction through the NI interface altered the objective intelligibility of the mix. As in other sections of this work (Chapters 5, 6 and 7), the mix was analysed utilising the objective computational metric termed the GP [65]. This measure gives an indication of the energetic masking which non-speech sounds will have on the speech.

Methodology of analysis

This section investigates both the overall change that the NI control has on the GP of the mix over the whole programme and the improvement the control gives on a scene by scene basis.

For the overall analysis, the GP was calculated at increments of 0.1 on the NI control, and averaged over the length of the programme. For the scene by scene analysis, the GP was calculated at the levels corresponding with the ACCESSIBLE, ENHANCED and TV MIX settings on the control. This allowed for a visualisation of the scenes where the greatest benefit was gained from the NI control.

The analyses in this section are estimations of the overall energetic masking potential of the non-speech objects. In addition to the inherent limitations of the GP, the mix had to be decomposed into the target (objects in the ESSENTIAL level, which were dialogue and any non-speech noises which could not be removed form the dialogue recordings) and masker (all remaining narrative importance levels). This provided a crude estimate of energetic masking and does not take into account the potential value of the narratively important

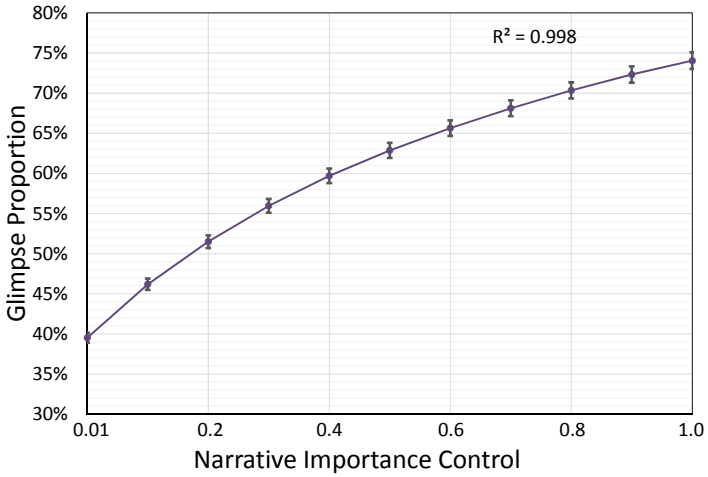


Figure 11.4 Glimpse proportion (GP)for dialogue compared with other audio objects, at 0.1 increments of the NI control, averaged over the length of the programme and plotted with standard deviation. This increase in GP across the range of the control shows the narrative importance approach can reduce energetic masking of dialogue.

sounds on overall ability to understand the speech or the masking value of background noises inextricable from the dialogue recording. Furthermore, as the GP in the form which it has been utilised in this work is a monaural measure, the GP for the left and right channel were calculated separately. These were then averaged to give an overall estimate.

Results

Figure 11.4 shows the average GP of the programme in 0.1 increments across the control. At the TV MIX end of the control it can be seen that the mix has a GP of 39.5%, which is almost doubled at the ACCESSIBLE end of the scale (74.0%). This indicates the energetic masking of the dialogue is effectively halved at the ACCESSIBLE end of the mix. This increase is not linear, with a second order polynomial providing the best fit (with a Pearson's R = 0.999), which is reflective of the logarithmic nature of the decibel scale. The error bars indicate one standard deviation. As these bars do not overlap for each increment from 0 to 0.8, this indicates that each increment of 0.1 offers a significant reduction of energetic masking. The final two increments are not significantly different.

Figure 11.5 shows the GP for each scene at three levels of the NI control: ACCESSIBLE, ENHANCED and TV MIX. The base level intelligibility of the original mix (TV MIX setting) is shown, with the additions of each of the ENHANCED and ACCESSIBLE settings shown on top of this value. As would be expected, there is a range of GP values from 0 in scenes where there is no dialogue, through to the scene with the highest base GP of 69.7% (scene #12 where Dylan diagnoses Magdelena in the main A&E area).

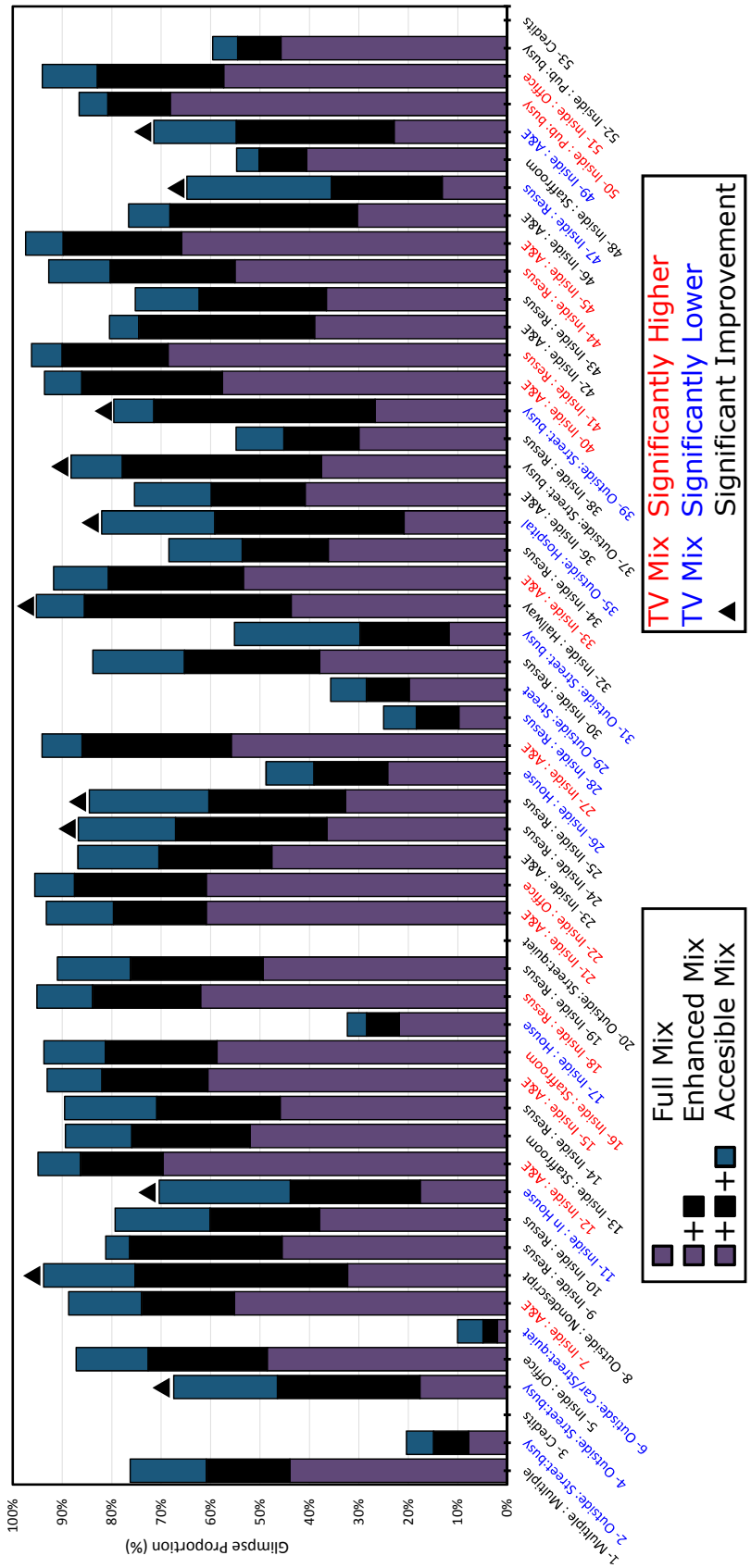


Figure 11.5 Glimpse proportion (GP) for each scene shown at the ACCESSIBLE, ENHANCED and ENHANCED settings of the NI control. Scenes which have significantly higher than average GP for the TV MIX are noted in red text and scenes with significantly lower GP are noted in blue text. Scenes which have significantly higher total improvement at the ACCESSIBLE setting, compared with the TV MIX setting are noted with a triangle.

A Chi Squared Goodness of Fit test was applied twice: first the TV MIX to determine whether the scenes in the original mix have equal energetic masking and secondly, to determine which scenes had a significant reduction in energetic masking due to the NI control (as defined by Eqn. 11.1). The scenes with no speech, scenes #3, #20 and #53, were not included in this analysis.

Base energetic masking in scenes The results of the first analysis showed that the GPs were not evenly distributed ($[z = 381, df = 49, p < 0.01]$). The upper and lower bounds of significance were calculated as $C_{sig} = 52.8\%$ and $C_{sig} = 27.9\%$ respectively, based on a two tailed test at the level $\alpha = 0.05$. This yielded 15 scenes which had significantly higher base GPs than others and 13 which have significantly lower GPs. These are noted in red and blue text in Figure 11.5 respectively.

Of the scenes with significantly higher GPs, all are indoor scenes, the majority of which have minimal action and, as such, have minimal non-speech sounds set at high levels. Of the scenes with significantly lower GPs, we can see that more than half are outdoor scenes. A further three of the scenes are #11, #17 and #26 where Danny is at home. In these scenes there is minimal dialogue and the music is high in the mix as it is conveying the majority of the narrative in these scenes. Of the remaining three scenes, one is scene #28 where Barbara crashes and Archie attempts to save her. Whilst this scene has more dialogue, it is a high emotion scene with many narratively important sounds high in the mix (see: Table 11.5).

It is also interesting to note that the expected value for GP was calculated to be 40.3%. The GP for scene #1, which describes previous events on 'Casualty' and combines excerpts from multiple episodes, had a base GP of 44.0%. The similarity of the expected GP for this episode as compared with the 'previously on' scene indicates that the audio mix characteristics of this episode are more likely to be broadly representative of recent 'Casualty' episodes.

Reduction in energetic masking The second significance test investigated for which scenes the ACCESSIBLE end of the scale provided a significant reduction in energetic masking, as compared with TV MIX. Total improvement was calculated as:

$$\text{Total Improvement} = GP_{\text{ACCESSIBLE}} - GP_{\text{TV MIX}} \tag{11.1}$$

Again a Chi Squared Goodness of Fit test was performed, this time on the total improvement between ACCESSIBLE and TV MIX (Eqn. 11.1). Again the three scenes with no speech were omitted. The improvement was seen to be significant ($[z = 218.5, df = 49, p < 0.01]$). The expected value was seen to be 36.4% indicating that, on average, the NI control could be expected to almost double the total GP value (given the expected value of 40.3% for the

GP of the TV MIX). This has the practical implication of halving the energetic masking in a scene. The upper bounds of significance were calculated as $C_{sig} = 48.25\%$, based on a one tailed test at the level $\alpha = 0.05$. A one tailed test was selected as it was only scenes where the control offered significant improvement which were of interest. Eleven scenes had significant improvements in GPs, with these being denoted in Figure 11.5 with small triangle above the corresponding column. Five of these scenes were outdoor scenes.

11.3.4 Discussion

This section has described the first integration of NI metadata assignment into a real-life production scenario. This process highlighted a number of key areas which could be improved to make the assignment of NI metadata in current production workflows feasible and has also reflected many of the themes from Chapter 10.

In this study, the process of assigning the metadata was cumbersome as it required reworking much of the existing Pro Tools template. Discussion with the Dubbing Mixer highlighted that consideration of the narrative importance concept in the track laying stage (which for 'Casualty' is undertaken by a separate individual, prior to the dub) would significantly streamline the process. Furthermore, the sentiment from many production staff in Study Three were echoed here with the dubbing mixer suggesting that it would not be likely to increase the track-laying time significantly, it would just require an alteration to the template used.

Another issue which caused the narrative importance mix to take longer than necessary to produce was the need to manually audition the different NI control playback levels, as the Pro-Tools compatible version of the auditioning plug-in was not ready. As the mix was primarily produced with the levels at the ACCESSIBLE end of the scale, this resulted in numerous routing errors being missed until the final mix was listened to during rendering. The ability to regularly and easily check the mix at different levels during production would have prevented this from happening. This reinforces that workflow integration must be supported by robust and intuitive tools.

The separation of the sync track into its dialogue, spot effect and atmosphere added time to the narrative importance mix process. It was also noted that, when set towards the ACCESSIBLE end of the NI control, some aspects of the recording such as clothing rustle became more obvious. These factors highlight challenges not only for the NI approach, but all object-based audio approaches which require greater separation of audio objects than current standard recording approaches can deliver. An additional challenge identified as part of this research, which applies to all object-based content, is the application of group compression.

The production process reaffirmed that the concept of ranking the audio objects by their importance to the narrative is quite a natural concept for post-production talent. Furthermore, as the narrative importance mix production progressed, the Dubbing Mixer became quicker at making the NI metadata assignments, indicating that with regular use the process could become quite natural for post production staff. Together with the results of Studies Two and Three, this suggests that a well-designed training programme for the NI metadata acquisition would streamline its integration into production workflows.

Finally, the analysis that was performed on the resulting narrative importance mix demonstrated the significant reduction in energetic masking of dialogue which the narrative importance process can produce. This was demonstrated by the proportion of energetic masking nearly halving between the TV MIX and ACCESSIBLE ends of the control. That such a large reduction could be achieved, whilst retaining the creative intent of the mix, indicates that the narrative importance approach provided the desired balance between improving accessibility (for those who find that background elements mask dialogue) and retaining narrative comprehension.

11.4 Public trial

The public trial launched 3 hours after the episode premiered on terrestrial television. It was hosted on BBC Taster: the BBC's public prototype platform [4] and was available on Firefox, Chrome and Safari browsers, and Android phones. It was not available on iPad or iPhone due to limitations in the WebAudio API on these browsers on iOS. The study was available for thirteen weeks, from 9/06/2019 to 8/09/2019, including two extensions that were requested by 'The Radio Times' and BBC's 'Points of View'. There were 8442 page views over the thirteen week, with 6228 of these being unique. 20.4% of these were on desktop, 67.1% on mobile and 13.4% on tablet. It was given a mean rating of 3.6 stars.

The landing page of the study included a description of the study titled 'Inside Story'. At the top of the page was a button labelled 'Try It'. When the button was pressed it brought up an iFrame containing the Standard Media Player and allowed the user to start interacting with the technology. Before the main episode content began, a short pre-roll video would play explaining the reason for the study and how to use the slider. When the content was first released this pre-roll video was a short instructional video written by the author and produced by BBC R&D and Casualty's Social Media manager. From the 12th June onwards, this was replaced with a longer video written by and staring the actress Gabriella Leon, who

[4] https://www.bbc.co.uk/taster/

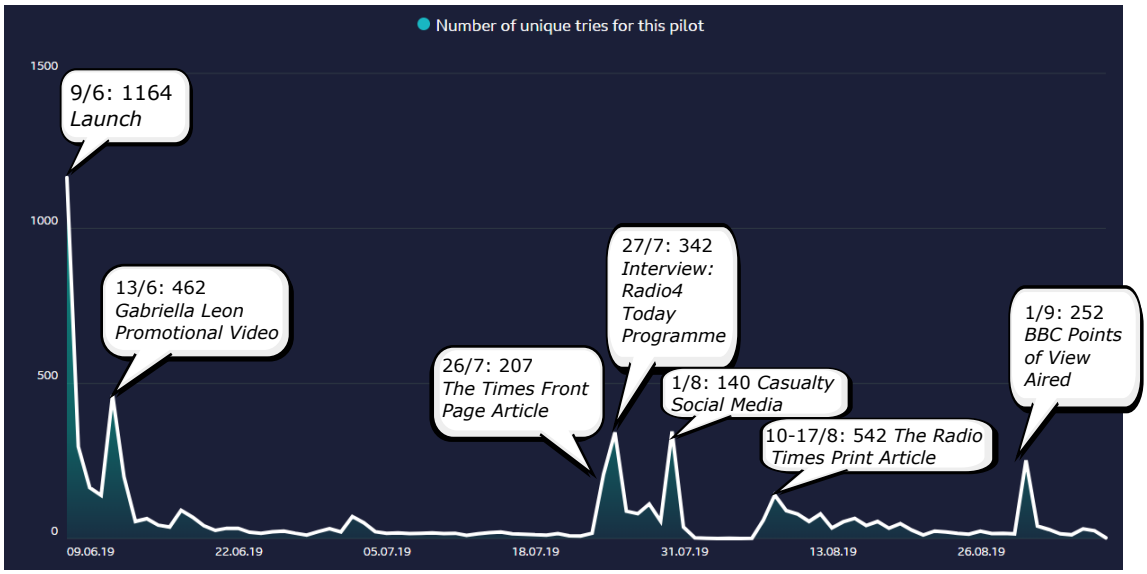


Figure 11.6 Number of unique visits to the Casualty Accessible and Enhanced Audio Taster page over the duration of the study, with the dates of key publicity events noted.

plays the character Jade in the episode. This included her own thoughts on the technology and was shot on the 'Casualty' set.

Publicity

The study was promoted through the University of Salford, BBC R&D and Casualty's social media channels. Hard of hearing charities, such as Action on Hearing Loss publicised the study, as did professional engineering societies including the IET. The study was also picked up by print, radio, TV and online media including 'The Times'[396], 'The Daily Mail' [397], 'BBC Radio4', 'The Radio Times' [398] and 'Points of View' [399]. These resulted in spikes in participation when these articles or programmes were distributed, which can be seen in Figure 11.6. When the instructional video with Gabriella Leon was released on Taster, a shorter promotional version of the video was published on Casualty's social media accounts on Twitter and Facebook. At the time of writing[5], these promotional videos had been viewed 1.9K and 42K times on Twitter and Facebook respectively. Figure 11.6 shows that participation was heavily influenced by publicity. Given this, we can assume that whilst a broad reach was likely achieved, hard of hearing individuals, older individuals and regular 'Casualty' viewers may be over-represented.

[5] 11/10/2019

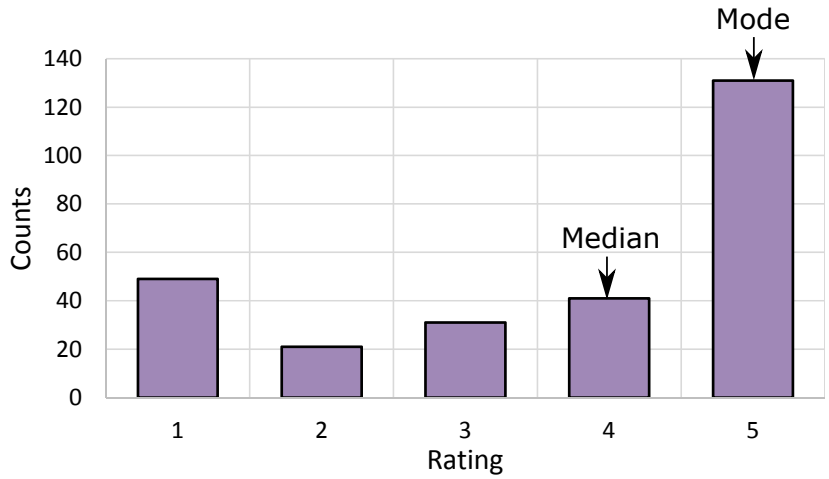


Figure 11.7 Number of participants who gave each overall rating for the Casualty Accessible and Enhanced Audio Trial, with the median and modal rating noted.

11.4.1 Feedback survey

Of those who participated in the study, 353 individuals rated the technology and 273 of those completed the feedback survey. This section describes the results of this feedback, along with preliminary discussions of the results. Only 7.3% of these responses came from the period prior to the instructional video being updated, thus was considered unlikely that the different instructional video would have had a significant effect the following results.

Rating

Those who rated the technology gave it a mean of 3.6 out of 5 stars. The distribution of ratings can be seen in Figure 11.7, along with the median rating which was 4 and the mode which was 5. There is a trend towards fewer people giving the pilot lower ratings, with the exception of the 1 star ratings. This may be influenced by those who interacted with the content on an iOS browser where, even though the content would not function, participants could still choose to rate the pilot.

Descriptive statistics

The results of the feedback survey can be seen in Table 11.6. This indicates the number of people who gave each response to each question, the percentage this represents of the total number who answered that question and the modal response (denoted in bold typeface). Of these, 273 individuals chose to answer at least one of the feedback questions and the response rate in Table 11.6 indicates the percentage of these who answered each question.

Table 11.6 Demographic questions on the BBC Taster page which participants could opt to answer, noting the possible multiple-choice responses, number of participants who gave each response (as a percentage of the total respondents to that question) and the response rate for that question. The modal response for each question is noted in bold typeface.

Demographic Questions	multiple-Choice Responses	Counts	
Are you D/deaf or hard of hearing? *Response Rate:* 96.3%	• Yes, D/deaf	14	5.3%
	• Yes, hard of hearing	105	39.9%
	• **No, neither**	**126**	**47.9%**
	• Not sure	18	6.8%
Do you usually watch Casualty? *Response Rate:* 88.3%	• Yes regularly	69	28.6%
	• Yes sometimes	28	11.6%
	• Yes rarely	28	11.6%
	• **Never**	**116**	**48.1%**
What age range are you?* *Response Rate:* 88.3%	• Under 16	6	2.5%
	• 16-24	21	8.7%
	• 25-34	20	8.3%
	• 35-44	15	6.2%
	• 45-55	34	14.1%
	• **55+**	**145**	**60.2%**
Are you male or female?* *Response Rate:* 89.4%	• Female	97	39.8%
	• **Male**	**140**	**57.4%**
	• Prefer not to say	7	2.9%
Where were you when you watched? *Response Rate:* 86.8%	• **Home living room**	**129**	**54.4%**
	• Other room at home	86	36.3%
	• Work	19	8.1%
	• Other	2	0.8%
	• Prefer not to say	1	0.4%

**denotes questions and their multiple-choice responses which are mandatory on the Taster platform*

Table 11.7 Feedback questions on the BBC Taster page which participants could opt to answer, noting the possible multiple-choice responses, number of participants who gave each response (as a percentage of the total respondents to that question) and the response rate for that question. The modal response for each question is noted in bold.

Feedback Questions	multiple-Choice Responses	Counts	
Did the audio control make a difference? *Response Rate:* 85.3%	• **Yes, lots**	**105**	**43.9%**
	• Yes a little	95	39.7%
	• None at all	14	10.4%
	• Not sure	14	5.9%
What difference did the control make? *Response Rate:* 87.5%	• **Easier to understand**	**101**	**43.3%**
	• More enjoyable	71	30.4%
	• Did not change	32	13.7%
	• Made it worse	10	4.3%
	• Not sure	19	8.2%
Should the BBC do more stuff like this?* *Response Rate:* 87.9%	• **Yes**	**222**	**92.5%**
	• No	12	5.0%
	• Not sure	6	2.5%

**denotes questions and their multiple-choice responses which are mandatory on the Taster platform*

It can be seen from Table 11.6 that the demographics from the feedback questions mirrors the expected demographics from Section 11.4, with older, hard of hearing and 'Casualty' viewers being over-represented. More than half of the participants, 60.2%, were over 55 years of age while 39.9% identified as hard of hearing and a further 5.3% identifying as D/deaf. The interactions between age and hearing loss are addressed in the following analysis. Half of the respondents, 51.9%, were viewers of 'Casualty' with half of those viewers (55.1%) being regular viewers. 39% identified as female and 58% as male.

As seen from the results, 90.7% watched at home, either in the living room (54.4%) or another room at home (36.3%). This indicates that the scenarios under which the study was watched were likely to be ecologically valid.

It is evident from Table 11.6 that participants found the control made a difference to their viewing experience and that that difference was overwhelmingly positive. 83.6% of respondents noticed the control having an effect, with 43.9% indicating the control made lots of difference. Almost three quarters of respondents indicated that the control had a positive effect, making the content easier to understand (43.3%) or more enjoyable (30.4%). The vast majority of respondents, 92.5% indicated 'Yes' in response to the question *'Should the BBC do more stuff like this?'*

Analysis

An analysis was undertaken to determine whether the known demographics of the participants had an impact on how they rated the pilot and the type and magnitude of the benefit they got from it. Only the age group, hearing loss and 'Casualty' viewership was considered in this analysis as the viewing location was skewed heavily to a single viewing location – the home – and as such, was deemed unlikely to reveal any significant differences. Similarly, due to the heavy skew in the feedback question of 'Should the BBC do more stuff like this?' this question was excluded from analysis. Gender was only included in the feedback survey as it was a mandatory question on the Taster platform, and was not considered pertinent to individuals' feedback. For this reason, it is not included in this analysis.

This analysis was undertaken using Pearson's Chi Squared test for independence of count data [270] and was implemented using the GMODELS package in R [400]. In order to yield a reliable Chi Squared value, all compared conditions must yield an expected count which is greater than one and less than 20% of the compared conditions must have an expected value which is less than five [270]. In order to meet these conditions, analysis was performed using 2 dimensional contingency tables, rather than higher dimensional tables containing more than two variables. Furthermore, some of the responses had to be grouped together to ensure these assumptions were not violated. For age group, the responses were grouped into

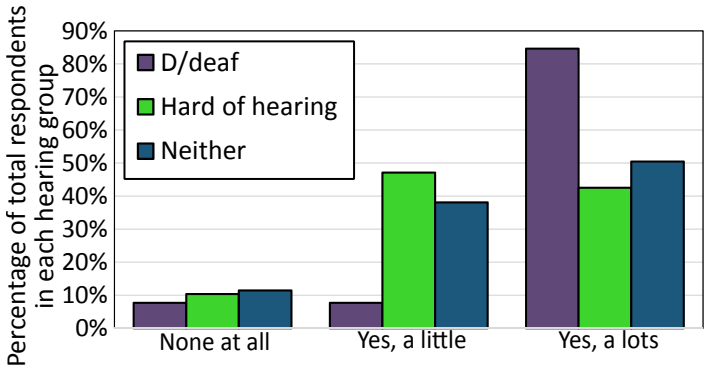


Figure 11.8 Responses to the question 'Did the audio control make a difference?' Responses are given as a percentage of respondents in each response category for the question: 'Are you D/deaf or hard of hearing?'

three new categories: Under 34, 35-54 and 55+. For 'Casualty' viewership the responses 'sometimes' and 'rarely' were grouped together as 'casual viewership'. All 'not sure' and 'skipped' responses were omitted.

Hearing loss Figure 11.9 shows the responses to the first two feedback questions decomposed by response to the question 'Do you identify as D/deaf or hard of hearing?' From Figure 11.8, it appears that those who identified as D/deaf were more likely to find the control made 'Lots' of difference. Inspecting what difference those who responded with 'Yes, lots' gave to the question 'What difference did the control make?' showed that 90% of responses are positive (either easier to understand, or more enjoyable). This indicates that those who identified as D/deaf, gained the most positive benefit. There are two main hypotheses for why this might be. For those making use of residual hearing for speech perception, attenuation of the MEDIUM IMPORTANCE and LOW IMPORTANCE sounds likely made them completely inaudible. For those relying primarily on subtitles, there would still be more information available than usual from the audio stream. This would include the temporal envelope of which better perception of would improve audio-visual congruency with the subtitles. Additionally. it would aid the perception of prosody and tone, which would complement the subtitles in conveying the emotion of the program.

Significance testing showed only borderline significance however for this relationship, with $[\chi^2 = 8.9, df = 4, p = 0.06]$ for the relationship between hearing loss and the difference observed and $[z = 1.82]$ for the relationship between 'Yes, lots' and 'D/deaf'. This may be due to the smaller sample population of D/deaf participants (n = 14) and suggests that further targeted investigations with this population are warranted.

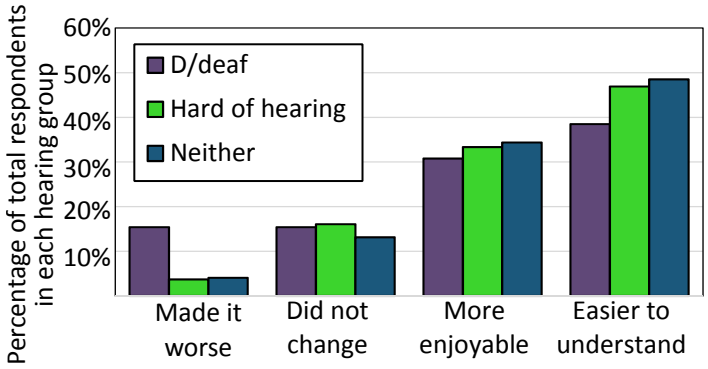


Figure 11.9 Responses to the question 'What difference did the control make?' Responses are given as a percentage of respondents in each response category for the question: 'Are you D/deaf or hard of hearing?'

The relationship between hearing loss and the type of difference the control made is seen in Figure 11.9. D/deaf listeners appear to be over-represented in the responses to 'Made it worse'. This, combined with the results above indicate that for those who likely have the highest levels of hearing loss, the effects of the control are more polarising: either giving greater benefit or greater detriment. Given the patterns observed in those with mild-moderate hearing loss in Chapter 8, this suggests that as hearing loss increases, the usefulness of non-speech sounds become increasingly polarised. Significance testing again showed no significant differences, but given the small sample of D/deaf listeners also warrants further investigation.

Age The relationship between age and the other variables was then explored, which showed significant differences for all variables, except for responses to the question 'How much difference did the control make?' Post-hoc testing was then conducted using standardised residuals to determine the significant pairs of responses.

For the relationship between age and rating the test statistic was $[\chi^2 = 31.5, df = 8, p < 0.001]$ Ratings decomposed by age group can be seen in Figure 11.10. Post-hoc testing using standardised residuals indicated that three pairs were significant, noted in Figure 11.10. There was a significant positive correlation between those Under 34 and ratings of 5 which was significant, with $[z = 1.98]$ and a significant negative correlation between those over 55 and ratings of 5 with $[z = -2.26]$, There was also a negative correlation between those under 34 and ratings of 2, with $[z = -1.98]$. There were two further borderline significant relationships for those aged 35-54: a negative correlations with ratings of 1 ($[z = -1.87]$) and a positive correlation with ratings of 5 ($[z = 1.94]$). Together these significant relationships indicate that younger viewers were more likely to rate the technology highly, whilst older

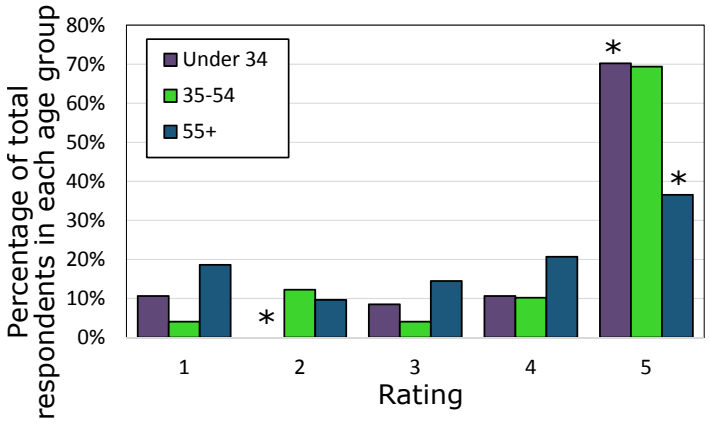


Figure 11.10 Participant ratings of the Casualty Accessible and Enhanced Audio trial. Responses are given as a percentage of respondents in each response category for the question: 'Are you D/deaf or hard of hearing?'

viewers were less likely to rate it highly. There are two likely hypotheses for why this is; familiarity of the interface and preference for personalisation. For younger viewers, who watch more content online or via iPlayer, the Standard Media Player interface was more likely to be familiar and subsequently they were able to use it with greater ease than their older counterparts. Alternatively, this result may indicate that regardless of hearing ability, younger viewers have a stronger preference for personalisation and customisable means for consuming their content.

The relationship between age and Casualty viewership had a test statistic of $[\chi^2 = 15.5, df = 4, p < 0.01]$ and showed that if someone under 34 responded they were likely to be a regular 'Casualty' viewer (z-score = 2.6). This is likely representative of the social media channels through which the study was publicised and indicates that much of the younger respondents were most likely recruited through the 'Casualty' social media channels. The relationship between age and hearing loss was also significant, with $[\chi^2 = 11.2, df = 4, p < 0.05]$. Interestingly the significant relationship was not between older listeners and identifying as hard of hearing, but between those identifying as D/deaf and those identifying as being Under 34 (z-score = 2.1). This likely points towards those with more severe hearing loss who are young being more likely to identify culturally with Deafness rather than being hard of hearing. Equally it may also be representative of the recruitment through Casualty's social media channels and through the promotion which Gabriella Leon gave to the project, as herself a young, Deaf advocate.

Significance testing between the individuals' age and response to 'What difference did the control make?' showed that those who said the control 'made it worse' were more likely to be Under 34 ($[\chi^2 = 18.1, p < 0.01]$). However, this only compared 5 responses in this

category to one in each of the older age groups respectively, so may be skewed by the small amount of responses of this type.

Casualty viewership The relationship between 'Casualty' viewership and the responses to the feedback questions was explored utilising the Chi Squared methodology above and shown not to have any significant relationships. This indicated that although a large proportion of the respondents were likely recruited through Casualty's social media channels, this did not measurably bias the results.

11.4.2 User interaction data

This section describes the data gathered from individuals' interactions with the media player, as described in Section 11.2.2. Over the period of the study, there were 5483 log files generated, of which 5028 were log files generated after midnight on the 13th June (after the fix for the logging routine was implemented). Given that there were 6366 individual page views from the 13th onwards, this indicates that 78.9% of those who visited the page interacted with the iFrame.

The aim of this section is to determine how many people interacted with the slider and the main content, to give an estimate of how long they interacted and whether there were any patterns in their usage behaviour.

Post-processing

In order to evaluate the interaction, the .json files had to be parsed and processed to identify any additional erroneous logs. The files were parsed utilising string operations in MATLAB.

Empty log files were first identified; these indicated that the individual did not or could not (due to iOS), interact with the main content. These were discarded leaving 1607 log files. This indicates that, of those who visited the page, 25.2% went on to interact with the main episode content.

Erroneous files were defined as those for which the only logged programme time was equal to exactly '0.0000'. These logs were considered erroneous as such a log could only be generated by interacting with the slider when the content was paused at the beginning and subsequently leaving the content. This resulted in a further 26 logs being discarded. The remaining 1581 logs are used in the ensuing analysis.

Interaction Time

As outlined in Section 11.2.2, the total spent interacting with the content cannot be calculated since a time stamp was not recorded when individuals first interacted with the iFrame. However, an estimate of how long individuals spent listening to the content and trialling the control could be made.

In place of an initial time stamp, the first log for 'program time' was used. This assumes that individuals did not skip any content before their first interaction, which results in likely overestimating the interaction time. Given that the median value for first interaction time (in the programme content) was 25.6 seconds, this over-estimation is likely to be of the order of half a minute.

The interaction time is defined using the following equation:

$$\text{Time}_{\text{Interact}} = pt(1) + \Big[T(end) - T(1)\Big] \tag{11.2}$$

where pt(1) is the first entry in the programme time log
and T(1) and T(end) are the first and final entries in the timestamp log respectively

From Eqn 11.2 it can be determined that the mean interaction time was 457.0 seconds and the median was 154.0. There were a small number of entries that had interaction times that were larger than the length of the content. This indicates that they paused the content and returned to it at a later time or alternatively re-watched segments of the content. These individuals were included in the calculation of the mean and median interaction times. The interaction times can then be used to determine and visualise the rate of engagement throughout the content. This is shown in Figure 11.11, calculated at one minute intervals.

From these calculations it can be seen that the majority of participants only interacted with the content for a few minutes. This was to be expected in this type of study [401]. However, there are many who watched longer stretches of content and likely interacted with it in a more ecologically valid way: i.e. as if they were watching a real piece of content. For this reason, segments of the following analysis will only consider those who watched for longer stretches of time.

Interaction behaviour

Two types of interaction values were logged; changes to the slider value and changes to the Standard Media Player's volume control. Whilst changes to the volume control would not trigger a new log, the change would be recorded when the slider was next altered.

The majority of individuals had the Standard Media Player's volume control set to either 0.7 or 1, with these values accounting for 52.4% and 40.1% of the unique volume values

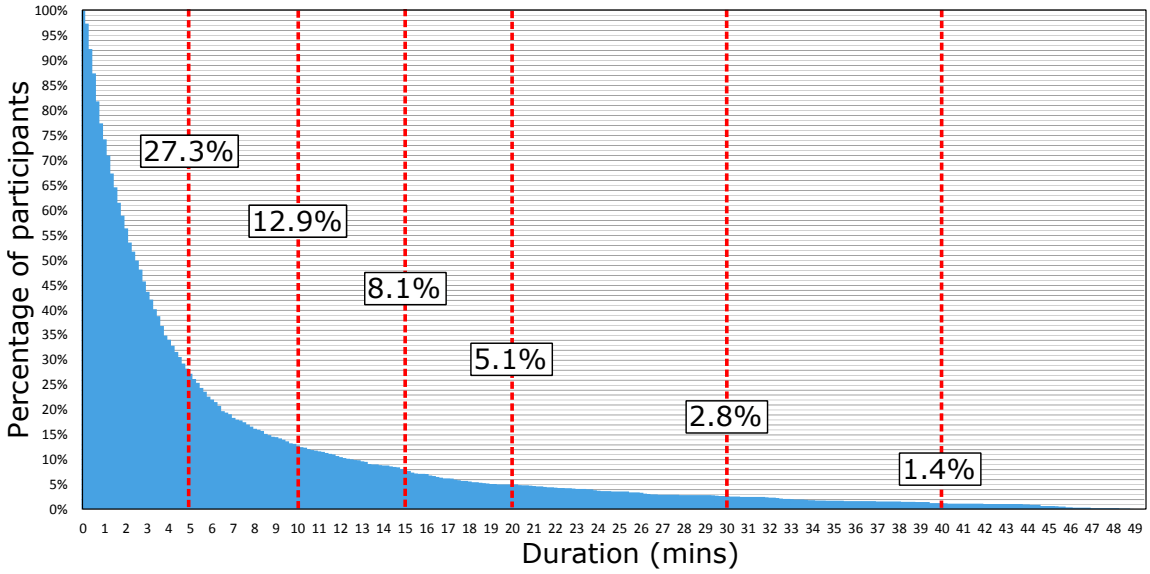


Figure 11.11 Estimated duration of interaction with the main content up to 49 minutes, shown at one minute intervals as a percentage of the total participants who engaged with the content. Percentage of total participants who engaged for a period of 5 mins, 10 mins, 15 mins, 30 mins, 30 mins and 40 mins are noted.

respectively. The default volume setting for the Standard Media Player is 0.7, indicating that the majority of participants did not change the volume control. A further 40.1% of participants increased the volume to the maximum during the pre-roll video. Only 117 (7.4%) participants adjusted the volume control during the main content, changing it an average of 2.3 times. Given this relative uniformity in the volume set by participants, this value was not investigated further.

The slider was set 24983 times or on average 15.8 times per participant. As these values were only logged once the participant released the slider, the number of different levels trialled across the slider range are likely to be significantly higher than this. To investigate the values which the slider was set, the histogram in Figure 11.12 was produced. The histogram has the same granularity as the control, with 100 increments between 0.01 and 1. As would be expected, the majority of values are set to either the TV MIX at one extreme or the ACCESSIBLE mix at the other extreme. These represent 19.5% and 20.1% of the total slider settings respectively. It can be seen that the remaining 60.4% of values were distributed relatively evenly.

To determine how the slider value was set in different segments of the content, each slider value and its corresponding programme time were plotted in a scatterplot. This can be seen in Figure 11.13. The time axis is delineated by scene, utilising the same scene definitions

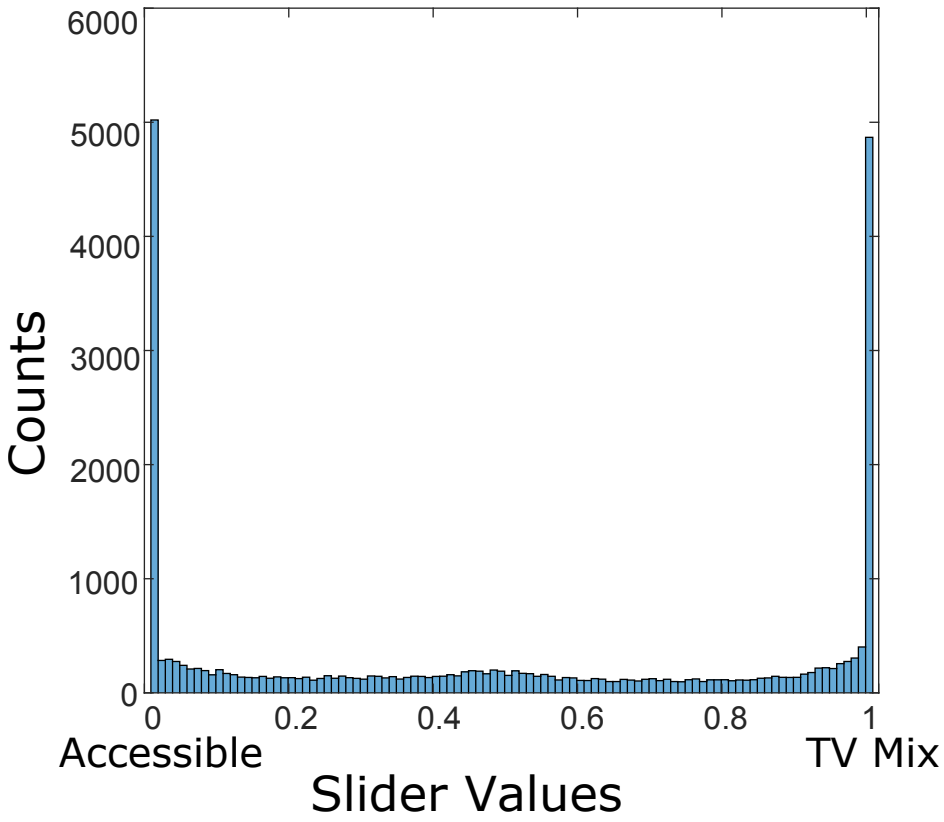


Figure 11.12 Histogram of the values the NI slider control was set to by participants during the main content, with 100 bins sized 0.01. The extremes of the control account for 19.5% and 20.1% at the TV MIX and ACCESSIBLE ends of the scale respectively.

as described in Tables 11.1 and 11.2. The slider value shown is representative of the value logged each time a participant changed the slider control.

We can see from Figure 11.13, that there is a much higher density of all values in the early scenes. Given that half of participants watched for less than three minutes this is to be expected. Similarly, it appears that those who selected one extreme or the other did so not just at the beginning, but consistently throughout the content. This may be the result of individuals skipping through the content and listening at the two extremes for comparison.

It can also be seen that there are a number of scenes beyond the 3 min mark which have a higher density of slider setting than the scenes surrounding them. To investigate whether these scenes had attracted significantly more changes in slider value than those around them, a Chi Squared Goodness of Fit test was used. Optimally, we would model how many people were left watching at each scene throughout the programme and see whether the number of changes was higher than expected for how many people remained. However due to the fact that interaction time was only an estimate, and participants could have skipped parts of the content, this was not possible.

As highlighted in Section 11.4.2, the median interaction time was approximately 457.0 seconds, corresponding with the end of scene #4. The behaviour of those who watched for a longer period, and thus likely interacted in a more ecologically valid manner, was analysed. This removed the densely populated initial portion where participants trialled many values quickly which was not indicative of long-term viewing behaviour.

As we cannot accurately model the drop off in interaction over time, the null hypothesis is taken as a uniform distribution where all scenes are equally likely to have a participant change the slider. Whilst this does not account for the drop off rate of participants, any caveats due to this assumption are considered in the results. Other caveats which will be considered were that scenes #8 and #9 were used regularly by Casualty Post-Production staff and BBC R&D staff as demonstration scenes. Furthermore, to take into account the different scene lengths, the average number of slider changes per second in each scene was used.

The methodology used to calculate the test statistics and the level of significant difference mirrors that used in Section 3.4.4. This yielded a test statistic of $[z = 348.3]$, allowing rejection of the null hypothesis at the level $[p = 0.01]$. Calculating the minimum and maximum level for significance at the level $\alpha = \pm 0.05$, giving a significance level of $C_{sig} > 10.3$ and $C_{sig} > 1.0$.

The results of this can be seen in Figure 11.14, which shows that there were 13 scenes that attracted significantly more changes in the slider value and 9 which attracted significantly less. Looking at the scenes which attracted more changes, two of the scenes, #8 and #9 were scenes which were known to be used by both BBC R&D staff and 'Casualty' staff as

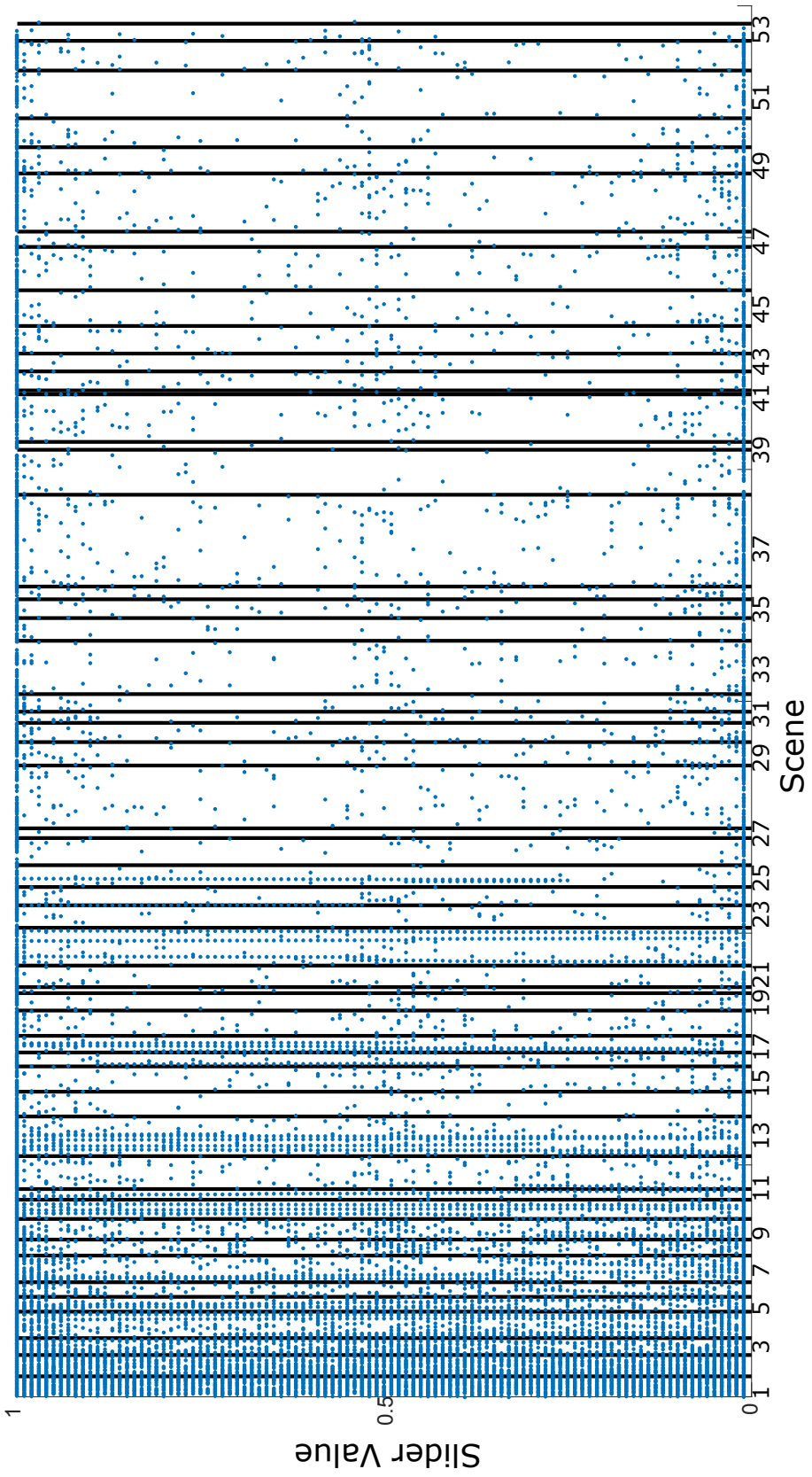


Figure 11.13 Scatterplot of the values the NI slider control was set to by participants during the main content against the time in the programme material the slider was changed, with scene boundaries shown. This shows that beyond the initial trials of the content there are a number of scenes where higher usage of the slider is apparent, particularly at intermediate values on the control.

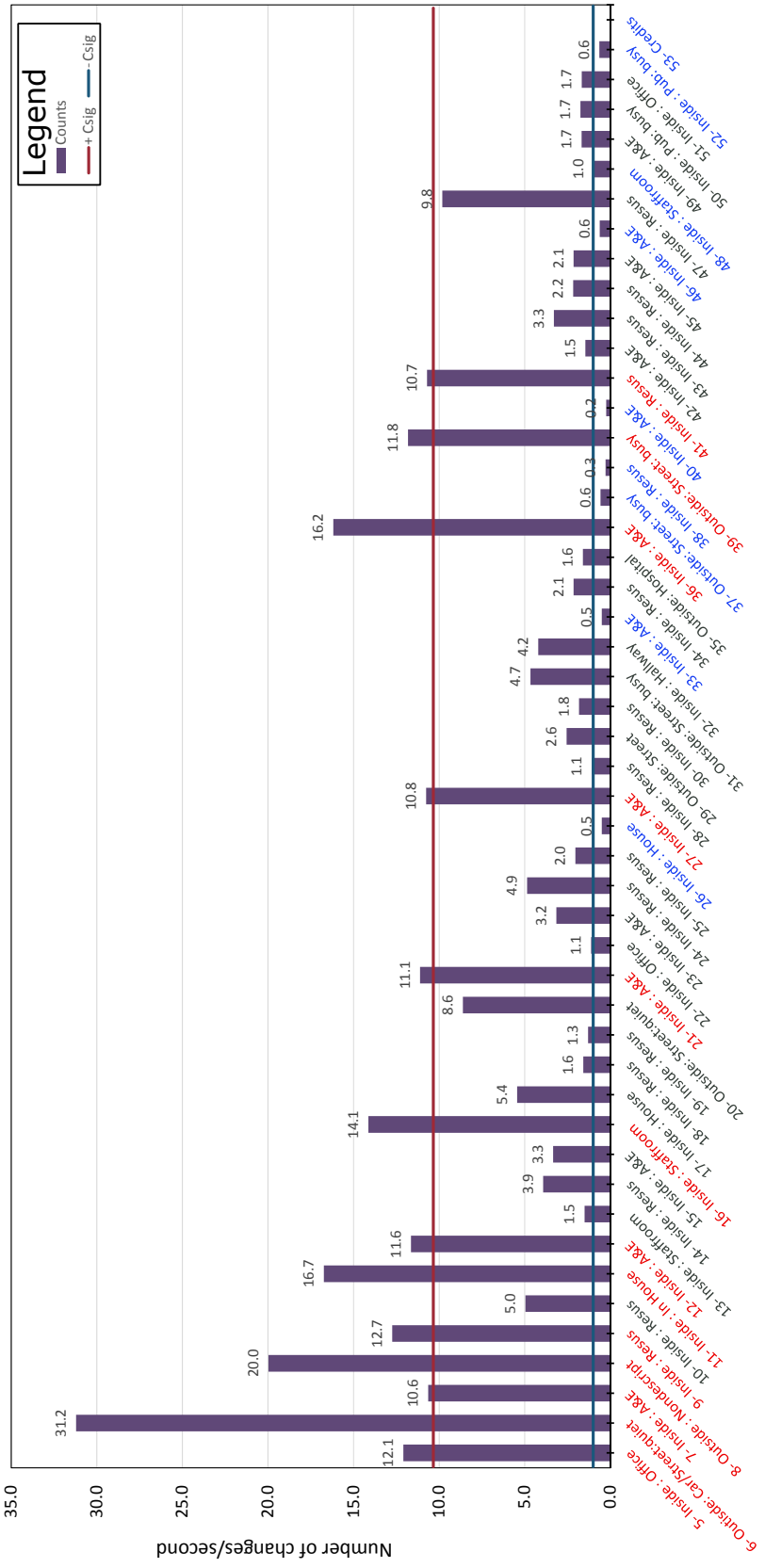


Figure 11.14 Average number of times the NI slider was changed per second, shown by scene, with the upper and lower bounds of significant difference noted in red and blue respectively. Those scenes which had significantly more changes per second are noted in red text and scenes with significantly less changes per second are noted in blue text.

example scenes. As a result, the higher amounts of changes in these scenes likely result from being used for demonstration purposes. Of the remaining 11 scenes, 9 were seen in Figure 11.5 to have either significantly higher or significantly lower base GP than the other scenes. Out of the 26 significantly different scenes (from scene #5 onwards), over one third of those attracted greater interaction. This suggests that, when encountering a scene which had significantly different energetic masking, individuals would adjust the control accordingly.

Together these results suggest that whilst the extremes of the scale were often used, as Figure 11.13 shows, participants explored many points on the scale. These results also suggest that the energetic masking of particular scenes may motivate interaction with the slider control. Though due to limitations in the collected data, there isn't sufficient evidence to make concrete conclusions about this.

11.4.3 Discussion

From the feedback data, it can be seen that the response to the study was overwhelmingly positive, and that this type of technology is something which audiences want to see more of. This positive response is despite a large proportion of those viewing on mobile and tablet (80.5%) likely using iOS, where the technology did not function correctly.

The percentage of those identifying as D/deaf was small, however their results suggest that the greatest benefit of the control may be experienced by D/deaf viewers. However, the results from this group also suggest that, for those with either more severe hearing loss or who utilise a signed language, the benefit (or detriment) of the technology is more polarised. Interestingly, the type and magnitude of the difference the control made was roughly even between normal and hard of hearing listeners. This indicates that the control offers personalisation ability which is valuable to audiences regardless of their hearing ability. Further to this, the technology appears particularly popular with younger viewers. This may be due to their familiarity with interactive content, their greater expectation for agency over their content consumption, or a combination of both.

It is also reassuring that regular viewership of 'Casualty' did not appear to have any affect on how participants rated the technology. This suggests both that those who were recruited through Casualty's social media channels did not bias the results and that addition of technology such as this was not viewed by the 'Casualty' fan-base as being detrimental to their enjoyment of the content.

The interaction data from users indicates that the study had a wide reach, with more than 1607 people interacting with the main content. Despite the limitations of the recorded interaction data, it can be concluded that participants explored not only the end points of the

NI control but also the intermediate values. This suggests that follow up controlled testing which compares a coarsely graduated scale with the finer grain scale used here is warranted.

The results from the interaction data also suggest that instances of significantly different energetic masking in content may motivate active engagement with the slider to alter the balance. Given, results from Strelcyk's work [15], which showed that those who reported greater problems understanding television content had greater remote control usage [15], this behaviour is plausible. Future investigation of this behaviour is warranted to determine whether viewers are willing to actively engage with content they deem not appropriate for their hearing needs.

This study has displayed the trade-off between large participant populations and the granularity of data obtainable. The limitations in the data highlight a number of particular areas where further, controlled investigation would be particularly illuminating. In particular, studies with a larger cohort of viewers who identify as D/deaf would be beneficial to determine the benefits of the control for this population and the mechanism of this benefit, in order to guide further system development. A more ecologically valid study, with multiple content pieces within the home would allow for greater investigation of individuals' behaviour when they deem the intelligibility of the content to be too poor (or too easy) for them. Further to this, a long-term study would indicate whether this behaviour is sustained with continued use or whether a single, compromise, setting is settled upon.

11.5 Results of the second research question

Part III has produced numerous findings addressing the second research question:

Can a system be designed to allow end-users to control the balance between audio objects for dramatic content which is simple to use and preserves comprehension?

This section summarises those findings, along with suggested directions for future work.

11.5.1 Narrative importance

In Armstrong's 2016 Whitepaper, he stated [21]:

If we are to be able to give our audience control over their listening experience then we need to be able to describe the items within a sound mix, their roles and how they can be manipulated.

This was the initial challenge which Part III had to address, in order to develop an effective end-user personalisation system. Taking inspiration from Gibson's theory of affordances [365] and its use in audio description [362], the concept of prioritising sounds based on their 'narrative importance' was developed. Chapter 9 proposed that narrative importance provided a common language between production staff and end-users. Using this narrative importance concept production staff could encode the value of audio objects within the object-based content's metadata. This encoding would then facilitate end-user manipulation of these objects, based on their importance. Study One in Chapter 9 and Studies Two and Three in Chapter 10 established that this hypothesised common language was in fact intelligible to both end-users and production staff.

Based on this language, prototype end-user and production tools were developed. For the end-user, the focus was on ensuring a simple interface to ensure a low barrier to access. Through the development of gain and attenuation laws for objects of each importance, it was possible to develop a personalisation system requiring only a single control. Integration of this control, in the form of a slider in the Standard Media Player in Chapter 11, showed that this system could naturally sit alongside existing controls for interacting with media content. Whilst there is still much to be explored around the user-experience of the NI control, this shows that simplicity of interface need not represent a trade-off between the personalisation control achievable and the integrity of the content.

11.5.2 User-experience of production staff

Study Two, Three and Four (Chapter 10 and 11) showed that acquisition of NI metadata in production workflows was feasible for many production staff. Familiarity with object-based workflows, practise and early integration into the production process were noted as essential to smooth integration into workflows. Widespread integration would require further research and development effort to be directed towards tool development and training resources.

Studies Two to Four highlighted many challenges which the entirety of object-based media must face: implementing group processing, specifically reverb and compression, and generating object-based metadata in live workflows. Of these, compression presents the most interesting challenge – one which speaks to the broader challenge of implementing loudness monitoring and control in object-based content. This work suggests the first step in resolving this challenge is investigation of how compression may be applied in the end-users rendering device. Such a change would prove disruptive to how content is currently produced. However, this change may also provide benefit, addressing the problems currently caused by cascaded compression (applied at production, transmission and, if viewed by a hearing aid user, at reception by the end-user).

The largest challenge faced by the development of production tools for narrative importance is that most content is not object-based. Whilst the push for the uptake of object-based audio is ongoing, hybrid tools which allow delivery of channel-based assets with object-based functionality are required. It is the opinion of the author that the development of hybrid tools hold the key for facilitating large and long-term trials of object-based functionality through channel-based delivery mechanism. It is likely that such trials are the only way in which broad uptake of object-based content will be achieved, by demonstrating both the desire for, and functionality of, object-based content for a broad range of audiences.

Finally, these studies have highlighted the importance of ensuring production staff are part of the metadata generation process. In the first instance, this manifested in comments which highlighted that the metadata acquisition process would be streamlined if tasks like track laying were completed with this output in mind. Going beyond this, the multitude of factors influencing the assignment of narrative importance seen in Chapter 10 demonstrate that is vital content creators are able to have control over the construction of their narrative. A poignant illustration of this is the importance assigned in Section 11.3.1 to non-speech sounds linking together over-arching series plots in 'Casualty'. Without the perspective of the production staff, such narrative elements would be lost.

11.5.3 End-user experience

Study One (Chapter 9) and Study Four (Chapter 11) established that the developed personalisation system is a positive step for access and agency in broadcast content consumption. These studies showed that the system achieved the aims of balancing enjoyment of content with ease of understanding, whilst retaining the *'depth'* and *'colour'* of the content. Study Four shows that these aims are achieved not only for those with hearing loss, but for broader audiences as well. This may be as a result of a general desire from audiences for greater agency over media, or may be indicative of the system addressing the media access needs of other audience segments. Either way, the developed system sets the foundation for future work exploring how more diverse media access needs can be met through object-based personalisation of broadcast content.

The studies here have yielded positive feedback on the narrative importance approach, however there is further work which should be undertaken to determine exact implementation detail for this approach. In particular, the results in Section 11.4.2 suggest that it is not only the end and middle points of the NI control which were used, but also other intermediate points. Determination of the degree of granularity which is required for implementing this tool is recommended at the next step in developing this approach. Furthermore, a quantitative comparative study which evaluates the benefit which may be gained from the narrative

importance approach, as compared with simple dialogue enhancement, should be undertaken. This work also suggests that a key unexplored direction for future work in this area is the interaction of accessible audio with subtitling services.

11.6 Part III Summary

This chapter describes the final of four user-experience studies in Part III. It has reported the first large-scale public trial of the narrative importance approach to accessible audio, called the 'Casualty Accessible and Enhanced (A&E) Audio Project'. Through collaboration with broadcasters and production teams, it has been possible address the research question:

Can a system be designed to allow end-users to control the balance between audio objects for dramatic content which is simple to use and preserves comprehension?

This study, along with the results in Chapters 9 and 10, has shown that the narrative importance approach to personalisation is valuable to audiences as well as being feasible in production workflows. In particular Part III has drawn the following conclusions:

- The narrative importance approach provides a common way of describing audio objects, which facilitates end-user personalisation.
- Acquisition of narrative importance metadata in production is feasible, but it should be performed once the content has already been nominally completed.
  - Integration into production workflows requires further development of tools, training resources and investigation of key challenges which compression and live-workflows present for object-based content.
- The narrative importance approach can almost halve the estimated energetic masking of dialogue in a mix whilst retaining the creative integrity the content.
- Audiences agree that the NI control achieves the aims of making content easier to understand and more enjoyable.
- The personalisation offered by the NI control is appealing to all audience segments, not only those with hearing loss, and particularly so to younger viewers.
  - Future investigation should explore of the implementation detail of the narrative importance approach and its integration with other access services.

Chapter 12

Summary and key contributions

This chapter outlines the novel contributions to knowledge this work has made. From these key contributions, three practical recommendations are made for translating this knowledge into the practise of developing technology and producing content for accessible audio applications. This chapter also presents a self-contained summary of the work completed, prior to the work's concluding remarks.

12.1 Contributions to knowledge

12.1.1 The audibility and accessibility problem space

The first key contribution to knowledge this work has made is the update of Armstrong's 'Audibility Problem Space' [21] and demonstrating the distinct challenges of creating *audible* and *accessible* content.

There has been an inconvenient tendency for the debate around intelligible broadcast content to conflate the words *audible* and *accessible* [396]. However, this work suggests that these terms are representative of two different challenges facing broadcasters. *Audibility* challenges, as defined by this work, refer to resolving issues which can degrade intelligibility for all listeners. When mumbled dialogue hits the headlines again [402], and again [403], and again [404], the issue which needs to be resolved is one of producing *audible* content. *Accessibility* challenges differ from this. It is defined by this work 'as the provision of services which ensure that any individual audience member's media access needs can be met and they can effectively engage with broadcast content'.

The delineation of these challenges was possible via projection of the identified intelligibility problems into a newly defined three dimensional space (Section 3.5). This space showed two dimensions were clearly dominated by the effects of individuals' sensory ac-

cess needs: *'General speech understanding'* and *'spatial/qualities of hearing'*. The third dimension however, showed challenges experienced across the surveyed cohort, irrespective of their access needs. Given the orthogonality of these dimensions, it was concluded that they were representative of two independent challenges: creating content which is *accessible*, i.e. meeting the media access needs of those with hearing loss, and creating content which is *audible*, i.e. content which is intelligible to the average viewer.

This gives a new lens with which to view the 'Audibility Problem Space'. For example, ensuring dialogue is recorded cleanly and not degraded by transmission codecs, is part of creating *audible* content. These are criteria which are either objectively met, or not: they are not dependent on the individual viewer. Balance between audio elements however, is highly subjective and the correct balance is different based on the needs of the viewer. This subjectivity makes it an *accessibility* challenge.

In addition to the delineation of *audibility* and *accessibility* challenges, the results of Chapter 3 highlighted updates required to Armstrong's 'Audibility Problem Space' in order to correctly reflect the current challenges. The update of the problem space can be seen in Table 12.1. Specifically three additions and one modification were made to Armstrong's 'Audibility Problem Space'. These are: MULTIPLE SPEAKERS AT A TIME, BALANCE BETWEEN AUDIO ELEMENTS and USER CONTROL OF REPRODUCTION. QUALITY AND AVAILABILITY OF CLOSED SUBTITLES needed to be modified to include the underlined text.

There are two requirements for intelligible broadcast content; first it must be *audible* then it must be *accessible*

Creating *audible* and *accessible* content are not only independent challenges, but they are sequential as well. This is to say that accessibility strategies should open up access to content, not compensate for poorly captured, produced and transmitted content. For example, personalised control of the audio balance, which is designed for accessibility, could be used to mitigate poorly recorded dialogue, but it should not be. Subtitling can be taken as another example. Personalisable subtitles with the ability to optimise the size and speed are superfluous if objective measures of quality, such as correct spelling, are not met. Accessibility strategies rely on the basis of *audible*, good quality content to function correctly for those who cannot access content without them.

In summary, *audibility* is the base requirement that content must first meet to be acceptable to the **average** viewer. *Accessibility* is the gold standard audible content should then aim for, to ensure **all** viewers have the ability to engage with it.

Table 12.1 Audibility and Accessibility Problem Space based on Armstrong's Audibility Problem Space [21] and the results of Chapter 3.

Audibility and Accessibility Problem Space				
1. Production /Direction	**2. Capture**	**3. Post Production**	**4. Broadcast**	**5. Home**
.1 Writing style – complexity of narrative	.1 Choice of microphone	.1 Voice processing, equalisation, noise reduction /gating, level control/limiting	.1 Output processing or other dynamic range control - not currently used on BBC TV	.1 Television sound quality – receiver, amplifier and speakers
.2 Direction and camerawork	.2 Microphone positioning	.2 Added music, level, type, purpose, style	.2 Audio encoding quality and any cascaded coding	.2 Use of mono, stereo or 5.1 loudspeakers
.3 Choice of shots – showing actor's face at vital moments in the narrative	.3 Skill of sound recordist	.3 Added sound effects, level, type purpose	.3 Quality and availability of closed subtitles	.3 Room Acoustics and the positions of the television and the viewer in the room
.4 Voices, accents, dialects & clarity of delivery	.4 Manual level control vs agc/limiter	.4 Added voiceover		.4 Background noise
.5 Choice of location - control of background noise	.5 Priority given to sound over video – e.g. microphone allowed in shot	.5 Video edit, choice of shots		.5 Viewer's hearing - level of interest in the programme - knowledge of the programme's subject matter - expectations of the programme - language skills - willingness to use subtitles
.6 Single Speaker at a time: direction, performance, style	.6 Retakes for sound	.6 Loudness control		.6 User control of reproduction
	.7 Voiceover booth matching to field recordings - ADR	.7 Dynamic range		
	.8 Use of digital compression and codec choice (mobile phone audio!)	.8 Stereo or 5.1 mix		
		.9 Inclusion of open subtitles		
		.10 Balance between audio elements		

12.1.2 Narrative Importance: Concepts and implementations

The second major contribution is the concept of 'Narrative Importance' and the associated production and end-user technology.

Previous approaches to accessible audio technology faced challenges caused either by over-simplifying the audio mix into binary categorisations (e.g. background-foreground) [140, 25, 147–149, 152] or presenting overly-complex interfaces with multiple variable parameters [170]. The problems which all these approaches had in common were the way in which audio objects were categorised: based on their physical, spectro-temporal characteristics.

This work's novel approach to the problem built on the fundamental of story-telling and the work of Gibson's 'Theory of Affordances' for visual perception. For *'the theory of affordances rescues us from the philosophical muddle of assuming fixed classes of objects, each defined by its common features and then given a name.'* Instead of a fixed class for an object, a sound effect for example, it can be defined by what it affords the viewer. Does the sound effect convey an important narrative point, like a gunshot in a *Whodunnit?* or is the sound effect just a part of the background atmosphere? By categorising objects based on what they afford the narrative, the challenge of grouping sounds and establishing their hierarchy is simplified. This idea is the foundation of the 'Narrative Importance' approach to accessible audio.

The affordance of any object is not absolute but taken with reference to the viewer. This work returned to the fundamentals of story-telling in television to determine the frame of reference for the affordances of audio objects. For in television content, *'[...] narrative is primary, and the audience member is actively engaged in constructing a narrative'* [355]. Not only do content creators use the affordance of sounds to construct the narrative in production, but as the audience, we shape these affordances based on our own experiences and the context in which we consume the content.

By returning to first principles of the narrative, we recognise that there are two key observers to the content: the content creator and the end-user. Whilst the building blocks they use to construct narrative meaning are the same, these blocks will afford each observer different things. Redundant non-speech objects can afford meaning and increase intelligibility (Chapter 5), but to others those same objects degrade intelligibility (Chapter 6 and 8). The same audio objects can be considered ESSENTIAL to some producers for conveying the narrative, whilst others consider them of little importance (Chapter 10).

The developed 'Narrative Importance' approach allows the perspectives of both production staff and end-user to be captured in the re-production. First, the production staff through assigning NI metadata to each object. Then the end-user through the 'Narrative

Importance' control, which allows them to adjust the mix based on their needs. By adopting this narrative importance approach, this work has been able to create a technology which retains the production staff's intended meaning, whilst accommodating what the content affords the viewer.

The narrative is important

Oral narratives are a fundamental human means of communicating information. This is particularly true of television which grew out of the oral story-telling traditions of radio (rather than the visual traditions of film) [405]. To ensure accessibility of television content, strategies cannot simply add functionality for the sake of it. They must build on the fundamentals of story telling.

For this reason, *accessibility cannot be an afterthought. It needs to be part of the creative process* [105]. The onus for this lies on both the technology developers and producers of content. Content producers must consider accessibility as they create. From inception, any developed accessibility strategies must consider the narrative form of the medium and how they can best integrate into the medium's creative process.

12.1.3 Accessibility is personal

The final contribution to knowledge is broadening the understanding of how hearing loss affects individual's media access needs and the usefulness of narratively important sounds. This work has demonstrated from the outset that the many and varied characteristics of hearing loss (summarised in Chapter 2) necessitate a personalisable approach to accessibility. This is further corroborated by the survey responses in Chapter 3, which demonstrated not only a variety of needs, but a desire from individuals to have greater control over audio reproduction.

The level of benefit or detriment gained from redundant non-speech objects is highly varied between individuals. This variability is in part due to the differing degrees to which those with hearing loss can utilise non-speech sounds. Part II has shown that the pure tone audiometric thresholds of an individual in their better hearing ear can, to some extent, predict this usefulness. This variability is also due to personal preference highlighted by Chapter 9's focus groups. Even with similar degrees of hearing impairment, preferences ranged from desiring speech only: *'if you lose the intonation [of the speech] the atmosphere is not very important, irrelevant'* to wanting a balanced mix with the background sounds *'because otherwise you're just listening to somebody reading a story'*. These preferences were not

even consistent for an individual: *'sat on the couch concentrating'* calls for a different mix than when splitting attention *'driving in a noisy Land Rover down a country lane'*.

This work and others [15] suggest that audiences are willing to exploit any agency available to improve their experience. Analysis of user interaction behaviour in Chapter 11 suggest that individuals will interact with the content more when the energetic masking of the dialogue changes. The results of Chapter 8 demonstrate that self-reported television speech understanding ability correlates with measures of speech perception ability in background noise with redundant non-speech audio objects present. Which is to say, audiences understand their own media access needs.

Finally, this agency over content is something which production staff are generally willing to provide. Chapter 10 shows us that at the heart of production staffs' considerations, is the audience, and that if providing personalisation increases access to content; *'that can only be a good thing'*.

Agency

These results highlight that one accessibility solution does not fit all and that agency over media is core to delivering truly accessible audio. This means providing as wide a variety of strategies as possible for interacting with content. As one focus group participant succinctly put it, when an accessibility strategy works it *'puts us back in control. We are in charge of what we hear and I think that's quite empowering.'*

12.2 Summary of work

This section provides a concise and self-contained summary of the key methodologies and results of each the three parts of this work.

12.2.1 Part I – The Questions

This doctoral work set out to improve the experience of broadcast audio for the 11 million individuals in the UK with some degree of hearing impairment [3] by exploiting new functionality in the next generation of broadcast audio technology. To achieve this, Part I set about identifying the most relevant and impactful research questions to address. In doing so, Chapter 2 first identified three dimensions of audio personalisation which could benefit end-users: *spatial separation*, *speech to background ratio* and *redundancy*. Using these dimensions, a systematic review of existing object-based audio personalisation research was

conducted, showing that the personalisation of redundant non-speech audio elements should be the key focus of this work.

Chapter 3 utilised a survey methodology to consult with end-users and establish the challenges individuals currently face accessing broadcast content. This built on existing studies and theory [15, 21, 102, 100, 251, 104], to update and extend our understanding of audience needs. The results of this survey first established the most difficult genres to understand speech in were drama and film. The level balance between speech and other audio objects was identified as a key challenge which audio personalisation could address. Chapter 3 also confirmed that end-users could not only benefit from control over audio reproduction, but also have a desire for greater agency over the way they watch content.

By bringing together the results of Part I, two research questions were developed and guided the remaining sections of this work. These were:

- What is the relationship between redundant non-speech audio objects and broadcast speech intelligibility, for normal and hard of hearing listeners?
- Can a system be designed to allow end-users to control the balance between audio objects for dramatic content which is simple to use and preserves comprehension?

12.2.2 Part II – The Science

Part II addresses the first of these questions through four perceptual studies. Chapter 4 begins this part by surveying the range of methodologies which exist for evaluating the intelligibility of speech, both in a broadcast context and more generally. It is concluded here that a mixed methods approach is required in addressing this research question, in order to generate results which are representative of the varied characteristics of hearing loss, repeatable and ecologically valid.

Chapter 5 presents Studies One and Two. These studies investigate the effect of redundant non-speech audio objects on keyword recognition in noise by normal hearing listeners. It has previously been shown that complementary cues, including semantic information, benefit intelligibility [319]. It is hypothesised that benefit is gained by using expectations and schemata of what an individual believes the audio object will contain to predict parts of the object for which no input signal is currently available [80]. These studies demonstrate that non-speech sounds can also offer normal hearing listeners complementary intelligibility cues. These cues can aid intelligibility in both low predictability speech (where there is no additional supporting context from the speech) and can build on speech-based context to further improve intelligibility in high predictability speech. However, the cue provided by

redundant non-speech audio objects have the potential to energetically mask speech, reducing their effectiveness if they are not presented separately (in the time domain) to the speech.

Chapter 6 presents Study Three, which begins to explore this research question with a cohort of hard of hearing individuals. Utilising a variation on the method developed in Chapter 4, this study shows that only some hard of hearing individuals experience improvement in keyword recognition in noise when redundant non-speech sounds are present. Furthermore, for some listeners, the presence of these redundant non-speech sounds can actively degrade keyword recognition. The degree to which a redundant non-speech sound will benefit, or degrade, keyword recognition in noise for hard of hearing listeners can be predicted by their average audiometric hearing threshold in their better hearing ear. Due to a number of limitations in Study Three, this could only be concluded for low predictability speech.

Chapter 7 and Chapter 8 build on this, through refinements to the methodology which allowed Study Four to show that audiometric thresholds would also predict the usefulness of redundant non-speech objects for high predictability speech. This effect is observed when other characteristics of an individuals' hearing loss are controlled for.

Through this methodological development, Chapter 7 provided a valuable resource for the broader research community. Chapter 7 describes the re-recording of the Revised Speech Perception in Noise test (termed the R^2SPIN) and its validation, as well as re-ordering of the test into a multiple speech to background paradigm. This work has also contributed an additional important resource in the form of the 'University of Salford media Accessibility and hearing Impairment Database' (termed USAID).

Part II has shown that individuals can use the complementary cues present in redundant non-speech audio objects to aid speech recognition in noise. Whilst this work has shown that the audiometric threshold in an individuals' better hearing ear predicts the benefit of redundant non-speech sounds, it is still not known which aspects of hearing or cognition cause the the varying degrees of benefit seen here. This suggests future work to undertake controlled investigations of the relationship between these varying degrees of benefit and an individual's stream segregation ability. Building on the results here, such a study may elucidate the mechanism behind the role of redundant non-speech audio objects in speech perception in noise by hard of hearing individuals. As such, Part II concludes, as all good science does, with further questions.

12.2.3 Part III – The Engineering

Part I showed that the current access service provisions are insufficient for many audience members and Part II showed that usefulness of non-speech audio objects varies between hard of hearing individuals. Part III aims to address these problems, through the development of a

technological solution which allows users to adjust the prominence and level of non-speech sounds to improve accessibility, based on how useful these sounds are to them.

To achieve this, Chapter 9 proposed a new approach to accessible audio called 'narrative importance'. Chapter 9 began by developing the philosophy behind the narrative importance concept, taking its inspiration from the established accessibility field of audio description [362] and Gibson's theory of affordances [365]. A prototype system, using this approach, was then evaluated in four qualitative user-experience studies with production staff and end-users.

The first of these studies aimed to establish the value of the narrative importance concept, and its implementation in the prototype system, for end-users. Using a focus group methodology, this study established that narrative importance and, the prototype system based on it, is considered by end-users as a viable direction for accessible broadcast audio. As in Part II though, no consensus exists on the benefit of key non-speech sounds. This again highlighted that personalisation is required to deliver accessible content.

Chapter 10 builds on this, establishing that the narrative importance concept was commensurate with production staff's prioritisation of audio objects in an mix. This was achieved through a case study with a single sound designer (Study Two) and a survey of a broad cohort of production staff (Study Three).

Study Two and Three demonstrated that, whilst this language is comprehensible across production staff, the manner in which it was implemented differs substantially. These studies also showed that acquisition of the necessary NI metadata during production workflows was feasible, but not without challenges. Studies One, Two and Three all highlighted that properly establishing the feasibility of a narrative importance based accessible audio personalisation system for broadcast would require a broader trial in real-life production environments and with ecologically valid audiovisual content.

Chapter 11 presented the *Casualty Accessible and Enhanced (A&E) Audio Trial*, the final study in Part III. This study evaluated the narrative importance concept in production workflows as well as conducting a large scale public trial of the end-user system. Through the participation of Casualty's post-production team, BBC R&D and thousands of end-users, this study has been able to gain an understanding of the value of end-user personalisation and also the production challenges it presents. This study showed not only an appetite for agency over media, but suggested that end-users were willing to be actively engaged in controlling their content. However, the degree of control which is desired is still unknown, and future work should investigate this. Furthermore, it demonstrated that those with the highest levels of hearing loss may have gained the greatest benefit from this approach to audio personalisation.

Finally, whilst workflow integration is possible, there are a number of outstanding challenges in the creation of production and delivery tools which need to be resolved.

These challenges notwithstanding, the four studies in Part III have together been able to answer the second question:

Can a system be designed to allow end-users to control the balance between audio objects for dramatic content which is simple to use and preserves comprehension?

And the answer is: **Yes.**

12.3 Concluding remarks

Sometimes it's necessary to go a long distance out of the way in order to come back a short distance correctly.

– Albee in "The Zoo Story" [406]

In 2016, at the commencement of this doctoral work, the implementation of object-based broadcast seemed a faraway prospect, potentially a decade away. Three years later, at the time of writing, content is being broadcast in numerous territories in object-based or object-based compatible formats. Crucial decisions about the implementation detail of next generation audio systems are no longer conjecture for broadcasters but agenda items for meetings. If the potential of object-based audio personalisation for accessibility is to be exploited, the time to do it is now.

Exploitation of this potential in a way that truly benefits the end-user whilst retaining the integrity of narrative content should be guided by three key tenets. First, *Accessibility is personal*. The selection and implementation of next generation audio codecs into broadcast systems should aim to maximise personalisation capabilities. Second, content must be both *audible* and *accessible*. The ability to personalise for accessibility should not be used as a crutch to compensate for poorly produced content. Finally, the *importance of narrative*. Narrative is the medium through which television conveys meaning and this fact must not only motivate the content we create, but the design of systems for consuming it too.

This work concludes where it began; in order to meet the needs of diverse audiences in a changing media landscape, access services must be made personalisable. However, by taking the circuitous rather than direct route, this work has been able to show that only by considering how we construct our own narratives can we ensure that the stories we tell are heard.

Bibliography

[1] S. Gregory and G. Hartley. *Constructing deafness*. Bloomsbury Publishing, 1990.

[2] T. Skelton and G. Valentine. 'It feels like being Deaf is normal': an exploration into the complexities of defining D/deafness and young D/deaf people's identities. *Canadian Geographer/Le Géographe Canadien*, 47(4):451–466, 2003.

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

[4] 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.

[5] 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. *Arch Intern Med*, 168(14):1522–1530, 2008.

[6] G. A. Gates and J. H. Mills. Presbycusis. *The Lancet*, 366(9491):1111–1120, 2005.

[7] Office for National Statistics. National population projections: 2014-based statistical bulletin, Oct, 2015. URL https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationprojections/bulletins/nationalpopulationprojections/2015-10-#older-people.

[8] Broadcasters Audience Research Board. Trends in television viewing 2016, Mar. 2017. URL http://www.barb.co.uk/download/?file=/wp-content/uploads/2017/03/BARB-Trends-in-Television-Viewing-2016.pdf.

[9] The Nielsen Company (US). the total audience report q1, 2017.

[10] OFCOM. Media Nations: UK 2019. Aug. 2019. URL https://www.ofcom.org.uk/__data/assets/pdf_file/0019/160714/media-nations-2019-uk-report.pdf.

[11] United Nations. Convention on the Rights of Persons with Disabilities (CRPD). *Resolution*, 61:106, 2016. URL 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.

[12] ITU - G3ICT. Digital inclusion: Making television accessible report, Nov. 2011. URL http://staging.itu.int/en/ITU-D/Digital-Inclusion/Persons-with-Disabilities/Documents/Making_TV_Accessible-English.pdf.

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

[14] Australian Government. Australian Broadcasting Corporation Act 1983: Compilation No. 28, 18 Mar. 2018. URL https://www.legislation.gov.au/Details/C2018C00079.

[15] O. Strelcyk and G. Singh. TV listening and hearing aids. *PLOS One*, 13(6), June 2018.

[16] D. Monzani, E. Galeazzi, G. M.and Genovese, A. Marrara, and A. Martini. Psychological profile and social behaviour of working adults with mild or moderate hearing loss. *Acta Otorhinolaryngologica Italica*, 28(2):61, 2008.

[17] you know what's cool? a billion hours.

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

[19] Standard KO-07.0127R1: TTA—Transmission and Reception for Terrestrial UHDTV Broadcasting Service, Revision 1, Dec. 2016.

[20] R. Brun. Successful demonstration of interactive audio streaming using mpeg-h audio at norwegian broadcaster nrk, July 2018. URL /www.audioblog.iis.fraunhofer.com/mpegh-nrk/.

[21] M. Armstrong. BBC White paper WHP 324: From Clean Audio to Object Based Broadcasting, Oct, 2016. URL http://www.bbc.co.uk/rd/publications/whitepaper324.

[22] J. Popp, M. Neuendorf, H. Fuchs, C. Forster, and A. Heuberger. Recent advances in broadcast audio coding. In *Proc. 9^{th} IEEE International Symposium on Broadband Multimedia Systems and Broadcasting*, pages 1–5, London, 2013. IEEE.

[23] J. Herre, H. Purnhagen, J. Koppens, O. Hellmuth, J. Engdegård, J. Hilper, L. Villemoes, L. Terentiv, C. Falch, A. Hölzer, et al. MPEG spatial audio object coding - the ISO/MPEG standard for efficient coding of interactive audio scenes. *Journal of the Audio Engineering Society*, 60(9):655–673, 2012.

[24] A. Silzle, R. Schmidt, W. Bleisteiner, N Epain, and M. Ragot. Quality of experience tests of an object-based radio reproduction app on a mobile device. *Journal of the Audio Engineering Society*, 67(7/8):568–583, 2019.

[25] B. G. Shirley and P. Kendrick. The clean audio project: Digital TV as assistive technology. *Technology and Disability*, 18(1):31–41, 2006.

[26] B. G. Shirley and P. Kendrick. ITC Clean Audio Project. In *Proc. 116th Audio Engineering Society Convention*. Audio Engineering Society, 2004.

[27] Senate Community Affairs References Committee et al. Hear us: inquiry into hearing health in Australia. May 2010. URL https://www.aph.gov.au/Parliamentary_Business/Committees/Senate/Community_Affairs/Completed_inquiries/2008-10/hearing_health/report/index.

[28] 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. *International Journal of Audiology*, 54(3):152–161, 2015.

[29] 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. *Occupational and Environmental Medicine*, 59(9):634–639, 2002.

[30] 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*, 46(5):455–463, 2002.

[31] M. Maassen, W. Babisch, K. D. Bachmann, H. Ising, G. Lehnert, P. Plath, P. Plinkert, E. Rebentisch, G. Schuschke, M. Spreng, et al. Ear damage caused by leisure noise. *Noise and Health*, 4(13):1, 2001.

[32] B. C. J. Moore. *Cochlear hearing loss: physiological, psychological and technical issues*. John Wiley & Sons, 2007.

[33] R. J. H. Smith, A. E. Shearer, M. S. Hildebrand, and G. Van Camp. *Deafness and hereditary hearing loss overview*. University of Washington, Seattle, 2014.

[34] P. M. Rabinowitz. Noise-induced hearing loss. *American family physician*, 61(9): 2759–2760, 2000.

[35] British Society of Audiology. Recommended procedure: pure tone air and bone conduction threshold audiometry with and without masking and determination of uncomfortable loudness levels. Sept. 2011. URL http://www.thebsa.org.uk/wp-content/uploads/2014/04/BSA_RP_PTA_FINAL_24Sept11_MinorAmend06Feb12.pdf.

[36] 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. *Journal of the American Academy of Audiology*, 23(10):779–788, 2012.

[37] M. Huckvale and G. Hilkhuysen. On the predictability of the intelligibility of speech to hearing impaired listeners. In *Proc. 1st International Workshop on Challenges in Hearing Assistive Technology*, Stockholm, Sweden, Aug. 2017.

[38] R. Plomp. Auditory handicap of hearing impairment and the limited benefit of hearing aids. *Journal of the Acoustical Society America*, 63(2):533–549, 1978.

[39] H. M. Bharadwaj, S. Masud, G. Mehraei, S. Verhulst, and B. G. Shinn-Cunningham. Individual differences reveal correlates of hidden hearing deficits. *Journal of Neuroscience*, 35(5):2161–2172, 2015.

[40] B. C. J. Moore and B. R. Glasberg. Simulation of the effects of loudness recruitment and threshold elevation on the intelligibility of speech in quiet and in a background of speech. *Journal of the Acoustical Society of America*, 94(4):2050–2062, 1993.

[41] C. W. Newman, J. A. Wharton, B. G. Shivapuja, and G. P. Jacobson. Relationships among psychoacoustic judgments, speech understanding ability and self-perceived handicap in tinnitus subjects. *Audiology*, 33(1):47–60, 1994.

[42] D. Baguley, D. McFerran, and D. Hall. Tinnitus. *The Lancet*, 382(9904):1600–1607, 2013.

[43] E. Villchur. Simulation of the effect of recruitment on loudness relationships in speech. *Journal of the Acoustical Society of America*, 56(5):1601–1611, 1974.

[44] K. Hopkins and B. C. J. Moore. The contribution of temporal fine structure to the intelligibility of speech in steady and modulated noise. *The Journal of the Acoustical Society of America*, 125(1):442–446, 2009.

[45] 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 a. *Journal of the Acoustical Society of America*, 129(2):852–863, 2011.

[46] D. P. Phillips and Michele M Carr. Disturbances of loudness perception. *Journal of the American Academy of Audiology*, 9:371–379, 1998.

[47] A. Davis and E. A. Rafaie. Epidemiology of tinnitus. In *Tinnitus handbook*, volume 1, page 23. San Diego, CA, 2000.

[48] L. McKenna, R. S. Hallam, and R. Hinchcliff. The prevalence of psychological disturbance in neuro-otology outpatients. *Clinical Otolaryngology & Allied Sciences*, 16(5):452–456, 1991.

[49] G. Cianfrone, D. Pentangelo, F. Cianfrone, F. Mazzei, R. Turchetta, M. P. Orlando, and G. Altissimi. Pharmacological drugs inducing ototoxicity, vestibular symptoms and tinnitus: a reasoned and updated guide. *European Review for Medical and Pharmacological Sciences*, 15(6):601–36, 2011.

[50] A. J. Heller. Classification and epidemiology of tinnitus. *Otolaryngologic Clinics of North America*, 36(2):239–248, 2003.

[51] A. Di Stadio, L. Dipietro, G. Ricci, A. Della Volpe, A. Minni, A. Greco, M. de Vincentiis, and M. Ralli. Hearing loss, tinnitus, hyperacusis, and diplacusis in professional musicians: A systematic review. *International journal of environmental research and public health*, 15(10):2120, 2018.

[52] M. B. Meikle and S. E. Griest. Asymmetry in tinnitus perceptions: Factors that may account for the higher prevalence of left-sided tinnitus. *Tinnitus*, 91:231–37, 1992.

[53] D. M. Baguley and D. J. McFerran. Hyperacusis and disorders of loudness perception. In *Textbook of tinnitus*, pages 13–23. Springer, 2011.

[54] H. Aazh, M. Knipper, A. A. Danesh, A. E. Cavanna, L. Andersson, J. Paulin, M. Schecklmann, M. Heinonen-Guzejev, and B. C. J. Moore. Insights from the third international conference on hyperacusis: Causes, evaluation, diagnosis, and treatment. *Noise & Health*, 20(95):162, 2018.

[55] U. Katzenell and S. Segal. Hyperacusis: review and clinical guidelines. *Otology & Neurotology*, 22(3):321–327, 2001.

[56] H. Davis and A. C. Goodman. Subtractive hearing loss, loudness recruitment, and decruitment. *The Journal of the Acoustical Society of America*, 38(5):922–923, 1965.

[57] S. Uppenkamp and M. Röhl. Human auditory neuroimaging of intensity and loudness. *Hearing Research*, 307:65–73, 2014.

[58] D. R. M. Langers, P. van Dijk, E. S. Schoenmaker, and W. H. Backes. fMRI activation in relation to sound intensity and loudness. *Neuroimage*, 35(2):709–718, 2007.

[59] K. Vermeire, A. Knoop, C. Boel, S. Auwers, L. Schenus, M. Talaveron-Rodriguez, C. De Boom, and M. De Sloovere. Speech recognition in noise by younger and older adults: Effects of age, hearing loss, and temporal resolution. *Annals of Otology, Rhinology & Laryngology*, 125(4):297–302, 2016.

[60] D. A. Nelson and R. L. Freyman. Temporal resolution in sensorineural hearing-impaired listeners. *The Journal of the Acoustical Society of America*, 81(3):709–720, 1987.

[61] J. R. Dubno, A. R. Horwitz, and J. B. Ahlstrom. Recovery from prior stimulation: Masking of speech by interrupted noise for younger and older adults with normal hearing. *The Journal of the Acoustical Society of America*, 113(4):2084–2094, 2003.

[62] P. J. Fitzgibbons and F. L. Wightman. Gap detection in normal and hearing-impaired listeners. *The Journal of the Acoustical Society of America*, 72(3):761–765, 1982.

[63] Y. Feng, S. Yin, M. Kiefte, and J. Wang. Temporal resolution in regions of normal hearing and speech perception in noise for adults with sloping high-frequency hearing loss. *Ear and Hearing*, 31(1):115–125, 2010.

[64] B. R. Glasberg, B. C. J. Moore, and S. P. Bacon. Gap detection and masking in hearing-impaired and normal-hearing subjects. *The Journal of the Acoustical Society of America*, 81(5):1546–1556, 1987.

[65] M. Cooke. A glimpsing model of speech perception in noise. *Journal of the Acoustical Society of America*, 119(3):1562–1573, 2006.

[66] A. J. Oxenham, J. G. W. Bernstein, and H. Penagos. Correct tonotopic representation is necessary for complex pitch perception. *Proceedings of the National Academy of Sciences*, 101(5):1421–1425, 2004.

[67] K. Hopkins, B. C. J. Moore, and M. A. Stone. Effects of moderate cochlear hearing loss on the ability to benefit from temporal fine structure information in speech. *The Journal of the Acoustical Society of America*, 123(2):1140–1153, 2008.

[68] H. Guest. Understanding and treating hidden hearing loss. URL https://www.actiononhearingloss.org.uk/finding-cures/our-biomedical-research/past-research-projects/understanding-and-treating-hidden-hearing-loss/.

[69] C. J. Plack, D. Barker, and G. Prendergast. Perceptual consequences of "hidden" hearing loss. *Trends in hearing*, 18, 2014.

[70] R. Schaette and D. McAlpine. Tinnitus with a normal audiogram: physiological evidence for hidden hearing loss and computational model. *Journal of Neuroscience*, 31(38):13452–13457, 2011.

[71] M. C. Liberman, M. J. Epstein, S. S. Cleveland, H. Wang, and S. F. Maison. Toward a differential diagnosis of hidden hearing loss in humans. *PloS one*, 11(9):e0162726, 2016.

[72] P. Z. Wu, L. D. Liberman, K. Bennett, V. De Gruttola, J. T. O'Malley, and M. C. Liberman. Primary neural degeneration in the human cochlea: evidence for hidden hearing loss in the aging ear. *Neuroscience*, 407:8–20, 2019.

[73] L. M. Viana, J. T. O'Malley, B. J. Burgess, D. D. Jones, C. A. C. P. Oliveira, F. Santos, S. N. Merchant, L. D. Liberman, and M. C. Liberman. Cochlear neuropathy in human presbycusis: Confocal analysis of hidden hearing loss in post-mortem tissue. *Hearing research*, 327:78–88, 2015.

[74] L. Fontan, J. Tardieu, P. Gaillard, V. Woisard, and R. Ruiz. Relationship between speech intelligibility and speech comprehension in babble noise. *Journal of Speech, Language, and Hearing Research*, 58(3):977–986, 2015.

[75] IEC60268-16. Standard, International Electrotechnical Commission, 2011.

[76] N. Miller. Measuring up to speech intelligibility. *Int. J. Lang. Comm. Disorders*, 48 (6):601–612, 2013.

[77] 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. *Journal of the Acoustical Society of America*, 61(5):1337–1351, 1977.

[78] 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 and Language*, 122(2):103–113, 2012.

[79] B. Lindblom. On the communication process: Speaker-listener interaction and the development of speech. *Augmentative and Alternative Communication*, 6(4):220–230, 1990.

[80] J.K. Bizley and Y.E. Cohen. The what, where and how of auditory-object perception. *Nature Reviews Neuroscience*, 14(10):693–707, 2013.

[81] I. J. Hirsh. The relation between localization and intelligibility. *Journal of the Acoustical Society of America*, 22(2):196–200, 1950.

[82] R.C. Bilger. *Speech recognition test development, In: E. Elkins ed. Speech recognition by the hearing impaired*, volume 14, pages 2–15. 1984.

[83] Z. Yang, J. Chen, Q. Huang, X. Wu, Y. Wu, B. A. Schneider, and L. Li. The effect of voice cuing on releasing chinese speech from informational masking. *Speech Communication*, 49(12):892–904, 2007.

[84] R. L. Freyman, U. Balakrishnan, and K. S. Helfer. Effect of number of masking talkers and auditory priming on informational masking in speech recognition. *Journal of the Acoustical Society of America*, 115(5):2246–2256, 2004.

[85] B. Spehar, S. Goebel, and N. Tye-Murray. Effects of context type on lipreading and listening performance and implications for sentence processing. *Journal of Speech, Language, and Hearing Research*, 58(3):1093–1102, 2015.

[86] 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. *Journal of Speech, Language, and Hearing Research*, 53(6):1629–1641, 2010.

[87] R. Ribback. Remote interpreting for live events & broadcast-inclusion in multiple ways. In *Proc. International Broadcast Convention Conference*. IET, 2015.

[88] OFCOM. Ofcom's code on television access services, Jan 2017. URL https://www.ofcom.org.uk/__data/assets/pdf_file/0020/97040/Access-service-code-Jan-2017.pdf.

[89] France CSA. Annual Report on the Accessibility of Television Programming for Persons with Disabilities and the Representation of Disability on the Air, 2017. URL https://www.epra.org/news_items/accessibility-of-disabled-persons-to-tv-programmes-and-disability-representation-on//-air-french-csa-report.

[90] European Regulators Group for Audiovisual Media Services. ERGA Special Task Report on the provision of greater accessibility to audiovisual media services for persons. Technical Report ERGA (2016) 12, 2016.

[91] Action on Hearing Loss. Progress on Pause: Spelling out the case for subtitle on on-demand services, 2015. URL https://www.actiononhearingloss.org.uk/how-we-help/information-and-resources/publications/research-reports/progress-on-pause-report/.

[92] UK Goverment. Digital economy act 2017, 2017. URL http://www.legislation.gov.uk/ukpga/2017/30/contents/enacted.

[93] Youtube Offical Blog. One billion captioned videos, Feb 2017. URL https://youtube.googleblog.com/2017/02/one-billion-captioned-videos.html.

[94] ITU Recommendation. ITU-R BS. 1770-4, Algorithms to measure audio programme loudness and true-peak audio level. 2015.

[95] EBU. Tech 3343 Guidelines for production of programmes in accordance with EBU R 128. Jan. 2016.

[96] ORF & 2DF ARD. TPRF-HDTV 2016 - Original sound for feature and documentary film. Technical report, Nov. 2016.

[97] BBC. Best practice guide sound mixing for BBC programmes. Technical report, 2018. URL http://dpp-assets.s3.amazonaws.com/wp-content/uploads/specs/bbc/TechnicalDeliveryStandardsBBCAudioMixGuidlines.pdf.

[98] Netflix. Netflix Sound Mix Specifications and Best Practices vOC-1-1. Technical report, July 2018.

[99] BBC. BBC Editorial Guidelines, "Hearing Impaired Audiences" . Technical report, 2011. URL http://downloads.bbc.co.uk/guidelines/editorialguidelines/pdfs/hearing-impaired.pdf.

[100] Royal National Institute for Deaf People. Annual survey report 2008, 2008.

[101] D. Cohen. Sound matters, BBC College of Production, Mar. 2011. URL http://www.bbc.co.uk/academy/production/article/art20130702112136134.

[102] Voice of the Listener and Viewer. VLV's Audibility of Speech on Television Project will make a real difference, June 2011. URL http://www.vlv.org.uk/documents/06.11PressreleasefromVLV-AudibilityProject-0800hrs1532011_002.pdf.

[103] Mike Armstrong. Audio processing and speech intelligibility: a literature review. In *BBC Research & Development Whitepaper*, 2011.

[104] P. Mapp. Intelligibility of Cinema & TV Sound Dialogue. In *Proc. 141st Audio Engineering Society Convention*, Sept. 2016.

[105] N. E. Youngblood, L. N. Tirumala, and R. A. Galvez. Accessible media: The need to prepare students for creating accessible content. *Journalism & Mass Communication Educator*, 73(3):334–345, 2018.

[106] M. Armstrong and M. Crabb. Exploring ways of meeting a wider range of access needs through object-based media - workshop. In *Conference on Accessibility in Film, Television and Interactive Media*, York, UK, Oct. 2017.

[107] M. Crabb, M. Heron, R. Jones, M. Armstrong, H. Reid, and A. Wilson. Developing accessible services: Understanding current knowledge and areas for future support. In *Proc. CHI Conference on Human Factors in Computing Systems*, page 216. ACM, 2019.

[108] C. Duarte and M. J. Fonseca. Multimedia accessibility. In *Web Accessibility*, pages 461–475. Springer, 2019.

[109] M. Oliver, B. Sapey, and P. Thomas. *Social Work with Disabled People*. Palgrave Macmillan, 2012.

[110] P. Romero-Fresco. In support of a wide notion of media accessibility: Access to content and access to creation. *Journal of Audiovisual Translation*, 1(1):187–204, 2018.

[111] T. R. Agus and C. Corrigan. Adapting audio mixes for hearing impairments. In *Proc. 3rd Workshop on Intelligent Music Production*, Sept 2018.

[112] H. Baumgartner, W. Hoeg, A. M. Koolwaay, J. Rennies, M. Wächtler, Schukrafft, W., and E. Bodenseh. Barrierefreie audiokommunikation – von der aufnahme bis ins wohnzimmer *ENGLISH: Barrier-free Audio Communication from On-site Recording to the Living Room*. In *28th Tonmeisteragung– VDT International Convention*, Cologne, Germany, Nov. 2014.

[113] A. Siegfried. Sprachverständlichkeit im fernsehton: Beschwerden, lösungen - *ENGLISH: Speech intelligibility: complaints, solutions*. In *Proc. 28th Tonmeisteragung– VDT International Convention*, Cologne, Germany, Nov. 2014.

[114] U. Krämer. Sprachverständlichkeit und hohe qualität – der schlüssel zu gutem fernsehton*ENGLISH: Speech intelligibility and high quality - the key to good TV sound*. In *Proc. 29th Tonmeisteragung– VDT International Convention*, Cologne, Germany, Nov. 2016.

[115] H Baumgartner, R. van Everdingen, B. Schreiner, M. Kahsnitz, U. Krämer, and A. Siegfried. Sprachverständlichkeit im fernsehen *ENGLISH: Speech Intelligibility in TV*. In *Proc. 29th Tonmeisteragung– VDT International Convention*, Cologne, Germany, Nov. 2016.

[116] T. Lund and E. Skovenborg. Loudness vs. speech normalization in film and drama for broadcast. In *Proc. Annual Technical Conference & Exhibition, SMPTE 2014*, pages 1–14. SMPTE, 2014.

[117] F. Andriessens. Sprachverständlichkeit um jeden preis? originalton bei spiel- und dokumentarfilm *ENGLISH: Speech intelligibility at any price? Original sound for feature and documentary film*. In *Proc. 30th Tonmeisteragung– VDT International Convention*, Cologne, Germany, Nov. 2018.

[118] E. Hildebrandt. *Sprachverständlichkeit im Fernsehen: Vorstellung von ausgewählten Teilaspekten zu diesem Thema im Kontext der Entwicklung einer Production Guideline ENGLISH: Speech intelligibility on television: Presentation of selected aspects of this topic in the context of the development of a production guideline*. PhD thesis, Institut für Komposition und Elektroakustik Universität für Musik und darstellende Kunst Wien, 2014.

[119] E. Erbert and E. Bodenseh. Sprachverständlichkeit im fernsehen: Arbeitsergebnisse und die daraus resultierende guideline *ENGLISH: Speech intelligibility in TV: A guideline*. In *Proc. 29th Tonmeisteragung– VDT International Convention*, Cologne, Germany, Nov. 2016.

[120] HBB4ALL: Hybrid Broadcast Broadband for All. Technical report, 2013-2016. URL https://cordis.europa.eu/project/rcn/191771/factsheet/en.

[121] T. Liebl and M. Weitnauer. Bessere sprachverständlichkeit im fernsehen, speziell für hörgeschädigte *ENGLISH: Improved speech intelligibility in television sound for hearing impaired people*. In *Proc. 29th Tonmeisteragung– VDT International Convention*, Cologne, Germany, Nov. 2016.

[122] M. Weitnauer, S. Goossens, and T. Liebl. Anpassbarer fernsehton für hörgeschädigte über hbbtv 2.0 *ENGLISH: Customizable TV sound for the hearing impaired via HbbTV 2.0*. In *Proc. DAGA*, Nüremberg, Germany, 2015.

[123] M. Bohne. *Untersuchungen zur Verbesserung der Sprachverständlichkeit für Hörgeschädigte im Rahmen des HbbTV-Standards ENGLISH: Improvement of speech intelligibility for hearing impaired people using the HbbTV-standard*. PhD thesis, Ostbayerische Technische Hochschule Amberg-Weiden, 2014.

[124] B.S.A. Recommended Procedure: Pure-tone air-conduction and bone conduction threshold audiometry with and without masking. Technical report, 2017.

[125] K. Ellis. Television's transition to the internet: Disability accessibility and broadband-based tv in australia. *Media International Australia*, 153(1):53–63, 2014.

[126] W. Hoeg and T. Lauterbach. *Digital audio broadcasting: principles and applications of DAB, DAB+ and DMB*. John Wiley & Sons, 2009.

[127] F. Utray, M. de Castro, L. Moreno, and B. Ruiz-Mezcua. Monitoring accessibility services in digital television. *International Journal of Digital Multimedia Broadcasting*, 2012, 2012.

[128] P. Looms. Making tv accessible in the 21st century. *Policy and marketing strategies for digital media.*, pages 43–59, 2014.

[129] Australian Communications Consumer Action Network (ACCAN), J. Slater, J. Lindstrom, and G. Astbrink. *Broadband solutions for consumers with disabilities*. Quest Publishing, 2010.

[130] ETSI. Digital Video Broadcasting (DVB); Specification for the use of Video and Audio Coding in Broadcasting Applications based on the MPEG-2 Transport Stream, Sept. 2009.

[131] C. D. Mathers. A study of sound balances for the hard of hearing. *NASA STI/Recon Technical Report N*, 91, 1991.

[132] A. R. Carmichael. Evaluating digital "on-line" background noise suppression: Clarifying television dialogue for older, hard-of-hearing viewers. *Neuropsychological Rehabilitation*, 14(1-2):241–249, 2004.

[133] A. Nakamura, N. Seiyama, A. Imai, T. Takagi, and E. Miyasaka. A new approach to compensate degeneration of speech intelligibility for elderly listeners-development of a portable real time speech rate conversion system. *IEEE Transactions on Broadcasting*, 42(3):285–293, 1996.

[134] E. Miyasaka, A. Nakamura, N. Seiyama, A. Imai, and T. Takagi. A new approach to compensate degeneration of hearing intelligibility for elderly listeners. In *Proc. 100th Audio Engineering Society Convention*. Audio Engineering Society, 1996.

[135] T. Komori, A Imai, N. Seiyama, R. Takou, T. Takagi, and Y. Oikawa. Study of TV sound level adjustment system for the elderly with speech rate conversion function. In *Proc. 137th Audio Engineering Society Convention*. Audio Engineering Society, 2014.

[136] A. Imai, T. Takagi, and H. Takeishi. Development of radio and television receiver with functions to assist hearing of elderly people. *IEEE Transactions on Consumer Electronics*, 51(1):268–272, 2005.

[137] T. Komori and T. Takagi. A system for adapting broadcast sound to the aural characteristics of elderly listeners. In *Proc. 121st Audio Engineering Society Convention*. Audio Engineering Society, 2006.

[138] T. Komori, A. Imai, N. Seiyama, R. Takou, T. Takagi, and Y. Oikawa. Development of a broadcast sound receiver for elderly persons. In *Proc. International Conference on Computers for Handicapped Persons*, pages 681–688. Springer, 2012.

[139] T. Komori, A Imai, N. Seiyama, R. Takou, T. Takagi, and Y. Oikawa. Development of volume balance adjustment device for voices and background sounds within programs for elderly people. In *Proc. 135th Audio Engineering Society Convention*, New York, U.S.A., 2013.

[140] J. Emmett. Dialogue enhancement for television sound. In *Proc. 104th Audio Engineering Society Convention*. Audio Engineering Society, 1998.

[141] 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. *Journal of the Audio Engineering Society*, 55(10):852–863, 2007.

[142] ETSI. ETSI TS101154 V2.4.1 Digital Video Broadcasting (DVB); Specification for the use of Video and Audio Coding in Broadcast and Broadband Applications, Feb 2018.

[143] EBU. EBU – TECH 3333: EBU HDTV Receiver Requirements, 2009.

[144] NorDig. Nordig unified requirements for integrated receiver decoders for use in cable satellite terrestrial and IP-based networks, 2009.

[145] Open IPTV Forum. OIPF Release 2 Specification Volume 2 - Media Formats, 2011.

[146] F. Rumsey. Hearing enhancement. *Journal of the Audio Engineering Society*, 57(5): 353–359, 2009.

[147] H. Müsch. Aging and sound perception: Desirable characteristics of entertainment audio for the elderly. In *Proc. 125th Audio Engineering Society Convention*. Audio Engineering Society, 2008.

[148] C. Uhle, O. Hellmuth, and J. Weigel. Speech enhancement of movie sound. In *Proc. 125th Audio Engineering Society Convention*. Audio Engineering Society, 2008.

[149] E. Vickers. Frequency-domain two-to three-channel upmix for center channel derivation and speech enhancement. In *Proc. 127th Audio Eng. Soc. Convention*. Audio Engineering Society, 2009.

[150] J. T. Geiger, P. Grosche, and Y. L. Parodi. Dialogue enhancement of stereo sound. In *Proc. 23rd European Signal Processing Conference (EUSIPCO)*, pages 869–873. IEEE, 2015.

[151] DTV4ALL: Digital Television for All. Technical report, 2013-2016. URL https://cordis.europa.eu/project/rcn/191846/factsheet/en.

[152] IRT, UAB, RAI, Brunel, TV Catalonia and RBB. Digital Television for All - 3.5 Emerging Access Services. 2010.

[153] TVC. D4.4 - Pilot-B Evaluations and recommendations. Technical report, 2016. URL http://pagines.uab.cat/hbb4all/sites/pagines.uab.cat.hbb4all/files/d4.4-tvc_pilot-b-evaluations-and-recommendations_v1.00.pdf.

[154] Virtual Dub. The "center cut" algorithm. URL http://www.virtualdub.org/blog/pivot/entry.php?id=102.

[155] M. Weitnauer and M. Bohne. Automatic and customizable improvement of the speech intelligibility from tv signals for hearing impaired people. In *Proc. 28th Tonmeisteragung– VDT International Convention*, Cologne, Germany, Nov. 2014.

[156] T. Liebl and M. Weitnauer. Bessere verständlichkeit im fernsehen *ENGLISH:* . In *Proc. 29th Tonmeisteragung– VDT International Convention*, Cologne, Germany, Nov. 2016.

[157] T. Liebl, M. Weitnauer, and M. Meier. Evaluierung eines ansatzes zur verbesserung der sprachverständlichkeit von stereosignalen im rundfunk *ENGLISH: Evaluation of an approach to improve speech intelligibility of stereo signals in broadcasting*. In *Proc. DAGA*, Aachen, Germany, 2016.

[158] IRT. D4.2 – Pilot-B Solution Integration and Trials. Technical Report D4.2, HBB4ALL. URL http://pagines.uab.cat/hbb4all/sites/pagines.uab.cat.hbb4all/files/d4.2-pilot-b-solution-integration-and-trials-2015.pdf.

[159] F. Matzura, S. Goossens, J. Groh, and E. Wilk. Verbesserung der sprachverständlichkeit von fertig gemischtem stereo-fernsehton *ENGLISH: Improved speech intelligibility of ready-mixed stereo TV sound*. In *Proc. DAGA*, Oldenburg, Germany, 2014.

[160] S. Goossens and T. Liebl. Verständlichkeit von fernsehton im tonkanalformat 3.0 – einfluss verschiedener abhörpegel auf das eingestellte sprache-hintergrund verhältnis. *ENGLISH: Intelligibility of TV sound in sound channel format 3.0 - Influence of different listening levels on the set speech to background ratio.* . In *Proc. 30th Tonmeisteragung– VDT International Convention*, Cologne, Germany, Nov. 2018.

[161] J. Paulus, M. Torcoli, C. Uhle, J. Herre, S. Disch, and H. Fuchs. Source separation for enabling dialogue enhancement in object-based broadcast with mpeg-h. *Journal of the Audio Engineering Society*, 67(7/8):510–521, 2019.

[162] H. Fuchs, S. Tuff, and C. Bustad. Dialogue enhancement - technology and experiments. *EBU Technical review*, 2, 2012.

[163] M. Mann, A. W. P. Churnside, A. Bonney, and F. Melchior. Object-based audio applied to football broadcasts. In *Proc. ACM International workshop on Immersive media experiences*, pages 13–16. ACM, 2013.

[164] C. Pike and Z. Watson. Virtual reality sound in the turning forest, May 2016. URL http://www.bbc.co.uk/rd/blog/2016-05-virtual-reality-sound-in-the-turning-forest.

[165] C. Baume. The mermaid's tears, Jan 2018. URL http://www.bbc.co.uk/rd/blog/2017-09-mermaids-tears-object-based-audio.

[166] I. Forrester and A. Churnside. The creation of a perceptive audio drama. In *Proc. NEM Summit*, 2012.

[167] J. Cox, M. Brooks, I. Forrester, and M. Armstrong. Moving object-based media production from one-off examples to scalable workflows. *SMPTE Motion Imaging Journal*, 127(4):32–37, 2018.

[168] 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 Transactions on Broadcasting*, 63(1):202–236, Mar. 2017. ISSN 0018-9316.

[169] 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 Journal*, 124(5):1–23, 2015.

[170] B. G. Shirley, M. Meadows, F. Malak, J. S. Woodcock, and A. Tidball. Personalized object-based audio for hearing impaired TV viewers. *Journal of the Audio Engineering Society*, 65(4):293–303, Apr. 2017.

[171] S. A. Silva. Object-based audio for televison production. In *Proceedings of the International Broadcasting Convention 2015*. IET, 2015.

[172] S. Coren. Most comfortable listening level as a function of age. *Ergonomics*, 37(7): 1269–1274, 1994.

[173] J. K. Willcox. Better tv sound for those with hearing loss. *Consumer Reports*, Mar. 2018. URL https://www.consumerreports.org/lcd-led-oled-tvs/better-tv-sound-for-those-with-hearing-loss/.

[174] 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. *Applied Science*, 8(1):59, 2018.

[175] T. Komori, T. Takagi, K. Kurozumi, and K. Murakawa. An investigation of audio balance for elderly listeners using loudness as the main parameter. In *Proc. 125th Audio Engineering Society Convention*. Audio Engineering Society, 2008.

[176] C. B. Pease. Combining the sone and phon scales. *Applied Acoustics*, 7(3):167–181, 1974.

[177] T. Liebl, S. Goossens, and G. Krump. Verbesserung der sprachverständlichkeit von fernsehton, speziell bei "voice-over-voice"-passagen (*ENGLISH: Improvement of "Voice-Over-Voice" speech intelligibility in television sound*). In *Proc. 28th Tonmeisteragung– VDT International Convention*, Cologne, Germany, Nov. 2014.

[178] M. Torcoli, A. Freke-Morin, J. Paulus, C. Simon, and B. G. Shirley. Background ducking to produce esthetically pleasing audio for tv with clear speech. In *Proc. 146th Audio Engineering Society Convention*. Audio Engineering Society, 2019.

[179] 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 Transactions on Broadcasting*, 2018.

[180] 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. In *Proc. 143rd Audio Engineering Society Convention*, New York, USA, Oct 2017.

[181] Y. Wu, E. Stangl, O. Chipara, S. Shabih Hasan, A. Welhaven, and J. Oleson. Characteristics of real-world signal to noise ratios and speech listening situations of older adults with mild to moderate hearing loss. *Ear and Hearing*, 39(2):293–304, 2018.

[182] R. Y. Litovsky. Spatial release from masking. *Acoustics Today*, 8(2):18–25, 2012.

[183] K. Belendiuk and R. A. Butler. Directional hearing under progressive impoverishment of binaural cues. *Sensory processes*, 2(1):58–70, 1978.

[184] T. L. Arbogast, C. R. Mason, and G. Kidd Jr. The effect of spatial separation on informational masking of speech in normal-hearing and hearing-impaired listeners. *Journal of the Acoustical Society of America*, 117(4):2169–2180, 2005.

[185] M. C. Heilemann, D. A. Anderson, and M. F. Bocko. Near-field object-based audio rendering on flat-panel displays. *Journal of the Audio Engineering Society*, 67(7/8): 531–539, 2019.

[186] T. Holman. *Sound for film and television*. Focal press, 2012.

[187] B. G. Shirley. *Improving Television sound for people with hearing impairments*. PhD thesis, University of Salford, 2013.

[188] 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. 2015.

[189] M. Pluymaekers, M. Ernestus, and R. Baayen. Articulatory planning is continuous and sensitive to informational redundancy. *Phonetica*, 62(2-4):146–159, 2005.

[190] 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. *Journal of American Academy of Audiology*, 23(8):590–605, 2012.

[191] D. J. Schum and L. J. Matthews. SPIN test performance of elderly hearing-impaired listeners. *Journal of American Academy of Audiology*, 3(5):303–307, 1992.

[192] 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. *Journal of Speech, Language, and Hearing Research*, 37(2):465–474, 1994.

[193] P. Souza, N. Gehani, R. Wright, and D. McCloy. The advantage of knowing the talker. *Journal of American Academy of Audiology*, 24(8):689–700, 2013.

[194] N. Hodoshima. Effects of urgent speech and preceding sounds on speech intelligibility in noisy and reverberant environments. In *Proc. 17th Annual Conference of International Speech Communication Association*, pages 1696–1699, San Francisco, USA, 2016. ISCA.

[195] 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 and Hearing*, 32(6):16–25, 2011.

[196] 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. *Journal of Educational Psychology*, 92(1):117, 2000.

[197] R. Barr, L. Shuck, K. Salerno, E. Atkinson, and D. L. Linebarger. Music interferes with learning from television during infancy. *Infant Child Development*, 19(3):313–331, 2010.

[198] M Aramaki, C. Marie, R. Kronland-Martinet, S. Ystad, and M. Besson. Sound categorization and conceptual priming for nonlinguistic and linguistic sounds. *Journal of Cognitive Neuroscience*, 22(11):2555–2569, 2010.

[199] M. Evans, T. Ferne, Z. Watson, F. Melchior, M. Brooks, P. Stenton, and I. Forrester. Creating object-based experiences in the real world. In *Proc. International Broadcast Convention Conference*. IET, 2016.

[200] FascinatE: Format-Agnostic SCript-based INterAcTive Experience. Technical report, 2010-2013. URL https://cordis.europa.eu/project/rcn/93759/factsheet/en.

[201] FascinatE: FP7/2007-2013, grant agreement no. 248138, 2010. URL www.fascinate-project.eu.

[202] B. G. Shirley, R. G. Oldfield, et al. Clean audio for tv broadcast: an object-based approach for hearing impaired viewers. *Journal of the Audio Engineering Society*, 63 (4):245–256, 2015.

[203] H. Fuchs and D. Oetting. Advanced clean audio solution: Dialogue enhancement. *SMPTE Motion Imaging Journal*, 123(5):23–27, July 2014.

[204] Orpheus: Object-based broadcasting – for European leadership in next generation audio experiences Fact Sheet. Technical report, 2015-2018. URL https://cordis.europa.eu/project/rcn/199837/factsheet/en.

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

[206] A. Silzle, M. Weitnauer, O. Warusfel, W. Bleisteiner, T. Herberger, N. Epain, B. Duval, N. Bogaards, C. Baume, U. Herzog, et al. "orpheus audio project: Piloting an end-to-end object-based audio broadcasting chain ". In *Proc. International Broadcast Convention Conference*, 2017.

[207] ICoSOLE: Immersive Coverage of Spatially Outspread Live Events. Technical report, 2013-2016. URL https://cordis.europa.eu/project/rcn/111011/factsheet/en.

[208] P. Golds, Br. Weir, S. Perrott, D. Evans, M. Paradis, P. Debenham, G. Thomas, H. Fraser, B. Gregory-Clarke, and S. Bason. Venue explorer. URL https://www.bbc.co.uk/rd/projects/venue-explorer.

[209] G. Kienast, W. Bailer, H. Fink, J. Schmidt, R. Bauwens, M. Wijnants, C. Pike, and R. Grandl. Deliverable 6.1: Initial demonstrators. Technical report, Oct. 2014. URL http://digital.joanneum.at/DOI/ICoSOLE-D6.1-JRS-InitialDemonstrators-v07.pdf.

[210] D. Marston, R. Bauwens, N. Peeters, and M. Matton. Deliverable 6.4: Second demonstration and field trial. Nov. 2016. URL http://digital.joanneum.at/DOI/ICoSOLE-D6.4-BBC-SecondFieldTrial-v09.pdf.

[211] 2-Immerse: Creating and Delivering Shared and Personalised Multi-Screen Broadcast and Broadband Experiences. Technical report, 2015-2018. URL https://cordis.europa.eu/project/rcn/199155/factsheet/en.

[212] J. Walker, D. L. Williams, I. C. Kegel, A. P. Gower, J. Jansen, M. Lomas, S. Fjellsten, UK Cisco, UK BT, NL CWI, et al. 2-immerse: A platform for production, delivery and orchestration of distributed media applications. In *Proc. International Broadcast Convention Conference*. IET, 2018.

[213] V. Vinayagamoorthy, R. Ramdhany, and M. Hammond. Enabling frame-accurate synchronised companion screen experiences. In *Proc. ACM International Conference on Interactive Experiences for TV and Online Video*, pages 83–92. ACM, 2016.

[214] A. Mason and M. Paradis. Adaptive, personalised "in browser" audio compression. In *Proc. 1st Web Audio Conference*. Citeseer, 2015.

[215] C Pike, P. Taylour, and F. Melchior. Delivering object-based 3d audio using the web audio api and the audio definition model. In *Proc. 1st Web Audio Conference*, 2015.

[216] T. Walton, M. Evans, D. Kirk, and F. Melchior. Exploring object-based content adaptation for mobile audio. *Personal Ubiquitous Comput.*, 22:707–720, Aug. 2018.

[217] T. Walton, M. Evans, D. Kirk, and F. Melchior. Does environmental noise influence preference of background-foreground audio balance? In *Proc. 141st Audio Engineering Society Convention*, Los Angeles, U.S.A., 2016. Audio Engineering Society.

[218] T. Walton, M. Evans, F. Melchior, and D. Kirk. Combining preference ratings with sensory profiling for the comparison of audio reproduction systems. In *142nd Audio Eng. Soc. Convention*. Audio Engineering Society, 2017.

[219] Principal Investigator: Prof. A Hilton. S3A: Future Spatial Audio for an Immersive Listener Experience at Home. Technical report, 2013-2019. URL https://gow.epsrc.ukri.org/NGBOViewGrant.aspx?GrantRef=EP/L000539/1.

[220] P. Demonte, Y. Tang, R. J. Hughes, T. Cox, B. Fazenda, and B. G. Shirley. Speech-to-screen: spatial separation of dialogue from noise towards improved speech intelligibility for the small screen. In *144th Audio Eng. Soc. Convention*. Audio Engineering Society, 2018.

[221] P. N. Tudor, P. J. Brightwell, and R. N. J. Wadge. Future models for live event broadcasting. In *Proc. International Broadcast Convention Conference*. IET, 2015.

[222] J. Francombe, J. Woodcock, R. J. Hughes, R. Mason, A. Franck, C. Pike, T. Brookes, W. J. Davies, P. J. B. Jackson, T. J Cox, et al. Qualitative evaluation of media device orchestration for immersive spatial audio reproduction. *Journal of the Audio Engineering Society*, 2018.

[223] A. Wilson, T. Cox, N. Zacharov, and C. Pike. Perceptual audio evaluation of media device orchestration using the multi-stimulus ideal profile method. In *Proc. 145th Audio Engineering Society Convention*. Audio Engineering Society, 2018.

[224] J. Francombe, J. Woodcock, R. J. Hughes, K. Hentschel, E. Whitmore, and T. Churnside. Producing audio drama content for an array of orchestrated personal devices. In *Proc. 145th Audio Engineering Society Convention*. Audio Engineering Society, 2018.

[225] J. Woodcock, J. Francombe, R. Hughes, R. Mason, W. J. Davies, and T. J. Cox. A quantitative evaluation of media device orchestration for immersive spatial audio reproduction. In *Proc. AES International Conference on Spatial Reproduction*. Audio Engineering Society, 2018.

[226] J. Francombe, R. Mason, P. J. B. Jackson, T. Brookes, R. Hughes, J. Woodcock, A. Franck, F. Melchior, and C. Pike. Media device orchestration for immersive spatial audio reproduction. In *Proceedings of the 12th International Audio Mostly Conference on Augmented and Participatory Sound and Music Experiences*, page 33. ACM, 2017.

[227] J. S. Woodcock, W. J. Davies, T. J. Cox, and F. Melchior. Categorization of broadcast audio objects in complex auditory scenes. *Journal of the Audio Engineering Society*, 2016.

[228] A. Franck, J. Francombe, J. Woodcock, R. Hughes, P. Coleman, D. Menzies, T. J. Cox, P. J. B. Jackson, and F. M. Fazi. A system architecture for semantically informed rendering of object-based audio. *Journal of the Audio Engineering Society*, 67(7/8): 498–509, 2019.

[229] M. F. Simon Galvez, 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)*, 23(11):1869–1878, 2015.

[230] M. Simon Galvez, I. Laghidze, L. Ward, A. Franck, B. G. Shirley, and F. Fazi. Multi-zone personalisation for hard of hearing listeners using object-based audio. In *Proc. 34th Reproduced Sound*. Institute of Acoustics, 2018.

[231] M. F. Simon Galvez, D. Menzies, R. Mason, and F. M. Fazi. Object-based audio reproduction using a listener-position adaptive stereo system. *Journal of the Audio Engineering Society*, 64(10):740–751, 2016.

[232] J. Jot, B. Smith, and J. Thompson. Dialog control and enhancement in object-based audio systems. In *Proc. 139th Audio Engineering Society Convention*. Audio Engineering Society, 2015.

[233] 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). In *Proc. 143rd Audio Engineering Society Convention*. Audio Engineering Society, 2015.

[234] H. Fuchs and D. Oetting. Advanced clean audio solution: Dialogue enhancement. In *Proc. International Broadcast Convention Conference*, Amsterdam, Netherlands, Sept. 2013. IET.

[235] A. Craciun, C. Uhle, and T. Bäckström. An evaluation of stereo speech enhancement methods for different audio-visual scenarios. In *Proc. 23rd European Signal Processing Conference (EUSIPCO)*, pages 2048–2052. IEEE, 2015.

[236] ImAc: Immersive Accessibility. Technical report, 2017-2020. URL https://cordis.europa.eu/project/rcn/211084/factsheet/en.

[237] T. Liebl and P. tho Pesch. D4.2-audio production tool. Technical report, Sept. 2018.

[238] A. Fidyka. Audio description in 360 degree videos: results from a focus group in barcelona. In *Proc. 9th International Symposium of Young Researchers*, 2018.

[239] D. Pearson. *Masters Thesis: Determining appropriate categorical boundaries for audio objects with regards to their importance to narrative clarity.* PhD thesis, University of Salford, 2017.

[240] M. Cooke, J. Barker, B. Shinn-Cunningham, and X. Shao. An audio-visual corpus for speech perception and automatic speech recognition. *Journal of the Acoustical Society of America*, 120(5):2421–2424, 2006.

[241] H. Levitt and L. R. Rabiner. Binaural release from masking for speech and gain in intelligibility. *Journal of the Acoustical Society of America*, 42(3):601–608, 1967.

[242] S. Sutojo, S. van de Par, and E. Schoenmaker. Contribution of Binaural Masking Release to Improved Speech Intelligibility for different Masker types. *European Journal of Neuroscience*, 2018.

[243] 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*, 38: 4–15, 1999.

[244] 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.

[245] P. Coleman, E. C. Cerón, J. Kim, C. Francombe, and J. Paulus. W03 - audio repurposing using source separation. In *Proc. 144th Audio Engineering Society Convention*, May 2018. URL http://www.aes.org/events/144/workshops/?ID=5917.

[246] M. Lopez and G. Kearney. Enhancing audio description: sound design, spatialisation and accessibility in film and television. In *Proc. 32nd Reproduced Sound*, Southampton, U.K., Nov. 2016. IOA.

[247] C. Baume. Even more or less: Designing a data-rich listening experience, Feb. 2019.

[248] D2.1 system architecture, 14 July 2017. URL https://2immerse.eu/wp-content/uploads/2018/01/d2.1_r2-system_architecture-clean.pdf.

[249] S. Gatehouse and W. Noble. The speech, spatial and qualities of hearing scale (SSQ). *International Journal of Audiology*, 43(2):85–99, 2004.

[250] W. Noble, N. S. Jensen, G. Naylor, N. Bhullar, and M. A. Akeroyd. A short form of the speech, spatial and qualities of hearing scale suitable for clinical use: The SSQ12. *International Journal of Audiology*, 52(6):409–412, 2013.

[251] Royal National Institute for Deaf People. Annual survey report 2005, 2005.

[252] A. Parbery-Clark, E. Skoe, C. Lam, and N. Kraus. Musician enhancement for speech-in-noise. *Ear and Hearing*, 30(6):653–661, 2009.

[253] M. Cohn, G. Zellou, and S. Barreda. The role of musical experience in the perceptual weighting of acoustic cues for the obstruent coda voicing contrast in American English. *Proc. 20th Annual Conference of International Speech Communication Association*, pages 2250–2254, 2019.

[254] S. M. K. Madsen, M. Marschall, T Dau, and A. J. Oxenham. Speech perception is similar for musicians and non-musicians across a wide range of conditions. *Scientific reports*, 9, 2019.

[255] M. A. Akeroyd, F. H. Guy, D. L. Harrison, and S. L. Suller. A factor analysis of the SSQ (speech, spatial, and qualities of hearing scale). *International Journal of Audiology*, 53(2):101–114, 2014.

[256] 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. *Frontiers in Ageing Neuroscience*, 6, 2014.

[257] S. Gatehouse and M. Akeroyd. Two-eared listening in dynamic situations: Audición con dos oídos en situaciones dinámicas. *International Journal of Audiology*, 45(sup1): 120–124, 2006.

[258] G. Singh and M. K. Pichora-Fuller. Older adults' performance on the speech, spatial, and qualities of hearing scale (SSQ): Test-retest reliability and a comparison of interview and self-administration methods. *International Journal of Audiology*, 49 (10):733–740, 2010.

[259] K. Demeester, V. Topsakal, J. Hendrickx, E. Fransen, L. van Laer, G. Van Camp, P. Van de Heyning, and A. Van Wieringen. Hearing disability measured by the speech, spatial, and qualities of hearing scale in clinically normal-hearing and hearing-impaired middle-aged persons, and disability screening by means of a reduced SSQ (the SSQ5). *Ear and Hearing*, 33(5):615–616, 2012.

[260] P. Zahorik and A. M. Rothpletz. Speech, spatial, and qualities of hearing scale (SSQ): Normative data from young, normal-hearing listeners. In *Proc. 167th Meetings on Acoustics*, volume 21, page 050007. ASA, 2014.

[261] J. Banh, G. Singh, and M. K. Pichora-Fuller. Age affects responses on the speech, spatial, and qualities of hearing scale (SSQ) by adults with minimal audiometric loss. *Journal of the American Academy of Audiology*, 23(2):81–91, 2012.

[262] A. Assarroudi, F. Heshmati Nabavi, M. R. Armat, A. Ebadi, and M. Vaismoradi. Directed qualitative content analysis: The description and elaboration of its underpinning methods and data analysis process. *Journal of Research in Nursing*, 23(1):42–55, 2018.

[263] H. Hsieh and S. E. Shannon. Three approaches to qualitative content analysis. *Qualitative health research*, 15(9):1277–1288, 2005.

[264] S. Elo and H. Kyngäs. The qualitative content analysis process. *Journal of advanced nursing*, 62(1):107–115, 2008.

[265] A. L. Chapman, M. Hadfield, and C. J. Chapman. Qualitative research in healthcare: an introduction to grounded theory using thematic analysis. *Journal of the Royal College of Physicians of Edinburgh*, 45(3):201–205, 2015.

[266] J. Francombe. *Perceptual evaluation of audio-on-audio interference in a personal sound zone system*. PhD thesis, University of Surrey (United Kingdom), 2014.

[267] I. Fellows. wordcloud. 2018. URL https://cran.r-project.org/web/packages/wordcloud/wordcloud.pdf.

[268] D. Meyer, K. Hornik, and I. Feinerer. Text mining infrastructure in r. *Journal of statistical software*, 25(5):1–54, 2008.

[269] J. Shlens. A tutorial on principal component analysis. *arXiv preprint arXiv:1404.1100*, 2014.

[270] A. Field, J. Miles, and Z. Field. *Discovering statistics using R*. Sage publications, 2012.

[271] H. F. Kaiser. The application of electronic computers to factor analysis. *Educational and psychological measurement*, 20(1):141–151, 1960.

[272] J. L. Horn. A rationale and test for the number of factors in factor analysis. *Psychometrika*, 30(2):179–185, 1965.

[273] J. Ruscio and B. Roche. Determining the number of factors to retain in an exploratory factor analysis using comparison data of known factorial structure. *Psychological assessment*, 24(2):282, 2012.

[274] W. R. Revelle. psych: Procedures for personality and psychological research. 2017.

[275] S. J. Van Wijngaarden, H. J. M. Steeneken, and T. Houtgast. Quantifying the intelligibility of speech in noise for non-native listeners. *Journal of the Acoustical Society of America*, 111(4):1906–1916, 2002.

[276] L. Ward and B. G. Shirley. Television dialogue; balancing audibility, attention and accessibility. In *Proc. Conference on Accessibility in Film, Television and Interactive Media*, 2017.

[277] W. Noble. Hearing, hearing impairment, and the audible world: A theoretical essay. *Audiology*, 22(4):325–338, 1983.

[278] Open Science Collaboration et al. Estimating the reproducibility of psychological science. *Science*, 349(6251):aac4716, 2015.

[279] 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. *The Journal of the Acoustical Society of America*, 95(2):1085–1099, 1994.

[280] S. Cameron and H. Dillon. The Listening in Spatialized Noise-Sentences Test (LISN-S): Comparison to The Prototype LISN and Results From Children With Either a Suspected (Central) Auditory Processing Disorder or a Confirmed Language Disorder. *Journal of the American Academy of Audiology*, 19(5):377–391, 2008.

[281] Etymotic. BKB-SIN^{TM} Speech-in-Noise test Version 1.03. 2005.

[282] J. Bench, Å. Kowal, and J. Bamford. the bkb (bamford-kowal-bench) sentence lists for partially-hearing children.

[283] HearCom. Matrix sentence test, 2010. URL http://hearcom.eu/prof/DiagnosingHearingLoss/AuditoryProfile/SpatialHearing.html.

[284] R. A. McArdle, R. H. Wilson, and C. A. Burks. Speech recognition in multitalker babble using digits, words, and sentences. *Journal of American Academy of Audiology*, 16(9):726–739, 2005.

[285] K. R. Duncan and N. L. Aarts. A comparison of the HINT and Quick Sin Tests. *Journal of Speech, Language Pathology and Audiology*, 30(2):86, 2006.

[286] N. Alghamdi, S. Maddock, R. Marxer, J. Barker, and G. J. Brown. A corpus of audio-visual lombard speech with frontal and profile views. *The Journal of the Acoustical Society of America*, 143(6):EL523–EL529, 2018.

[287] N. Tye-Murray, S. Hale, B. Spehar, J. Myerson, and M. S. Sommers. Lipreading in school-age children: the roles of age, hearing status, and cognitive ability. *Journal of Speech, Language, and Hearing Research*, 57(2):556–565, 2014.

[288] Y. Tang, M. Cooke, and C. Valentini-Botinhao. Evaluating the predictions of objective intelligibility metrics for modified and synthetic speech. *Computer Speech & Language*, 35:73–92, 2016.

[289] H. Fletcher. An empirical theory of telephone quality. *AT&T Internal Memorandum*, 101(6), 1921.

[290] ANSI S3.5. ANSI S3.5-1997 Methods for the calculation of the Speech Intelligibility Index, 1997.

[291] I. Holube and B. Kollmeier. Speech intelligibility prediction in hearing-impaired listeners based on a psychoacoustically motivated perception model. *Journal of the Acoustical Society of America*, 100(3):1703–1716, 1996.

[292] G. Carter, C. Knapp, and A. Nuttall. Estimation of the magnitude-squared coherence function via overlapped fast fourier transform processing. *IEEE transactions on audio and electroacoustics*, 21(4):337–344, 1973.

[293] J. M. Kates and K. H. Arehart. Coherence and the speech intelligibility index. *Journal of the Acoustical Society of America*, 117(4):2224–2237, 2005.

[294] F. Chen, O. Hazrati, and P. C. Loizou. Predicting the intelligibility of reverberant speech for cochlear implant listeners with a non-intrusive intelligibility measure. *Biomedical signal processing and control*, 8(3):311–314, 2013.

[295] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen. A short-time objective intelligibility measure for time-frequency weighted noisy speech. In *Proc. IEEE Int. Conference on Acoustics Speech and Signal Processing (ICASSP)*, pages 4214–4217. IEEE, 2010.

[296] H. J. M. Steeneken and T. Houtgast. A physical method for measuring speech-transmission quality. *Journal of the Acoustical Society of America*, 67(1):318–326, 1980.

[297] Ingrid M Noordhoek and Rob Drullman. Effect of reducing temporal intensity modulations on sentence intelligibility. *The Journal of the Acoustical Society of America*, 101(1):498–502, 1997.

[298] T. H. Falk, C. Zheng, and W. Chan. A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(7):1766–1774, 2010.

[299] B. Kollmeier, M. R. Schädler, A. Warzybok, B. T. Meyer, and T. Brand. Sentence recognition prediction for hearing-impaired listeners in stationary and fluctuation noise with FADE: Empowering the Attenuation and Distortion concept by Plomp with a quantitative processing model. *Trends in hearing*, 20:2331216516655795, 2016.

[300] Lionel Fontan, Isabelle Ferrané, Jérôme Farinas, Julien Pinquier, Julien Tardieu, Cynthia Magnen, Pascal Gaillard, Xavier Aumont, and Christian Füllgrabe. Automatic speech recognition predicts speech intelligibility and comprehension for listeners with simulated age-related hearing loss. *Journal of Speech, Language, and Hearing Research*, pages 1–12, 2017.

[301] M. Karbasi, A. H. Abdelaziz, H. Meutzner, and D. Kolossa. Blind non-intrusive speech intelligibility prediction using twin-hmms. In *Proc. 17th Annual Conference of International Speech Communication Association*, pages 625–629, 2016.

[302] R. Huber, H. Baumgartner, N. Moritz, and S. Goetze. Automatische überwachung der sprachverständlichkeit im rundfunkmaterial *ENGLISH: Automatic monitoring of speech intelligibility in broadcast material*. In *Proc. 30th Tonmeisteragung– VDT International Convention*, Cologne, Germany, Nov. 2018.

[303] C. Spille and B. T. Meyer. Listening in the dips: Comparing relevant features for speech recognition in humans and machines. *Proc. 18th Annual Conference of International Speech Communication Association*, pages 2968–2972, 2017.

[304] M. R. Schädler, A. Warzybok, S. D. Ewert, and B. Kollmeier. A simulation framework for auditory discrimination experiments: Revealing the importance of across-frequency processing in speech perception. *Journal of the Acoustical Society of America*, 139(5): 2708–2722, 2016.

[305] L. E. Humes, D. D. Dirks, T. S. Bell, C. Ahlstbom, and G. E. Kincaid. Application of the articulation index and the speech transmission index to the recognition of speech by normal-hearing and hearing-impaired listeners. *Journal of Speech, Language, and Hearing Research*, 29(4):447–462, 1986.

[306] K. S. Rhebergen, N. J. Versfeld, and W. A. Dreschler. Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise. *Journal of the Acoustical Society of America*, 120(6):3988–3997, 2006.

[307] R. M. Meyer and T. Brand. Comparison of different short-term speech intelligibility index procedures in fluctuating noise for listeners with normal and impaired hearing. *Acta acustica united with Acustica*, 99(3):442–456, 2013.

[308] J. M. Kates and K. H. Arehart. The hearing-aid speech quality index (HASQI) version 2. *Journal of the Audio Engineering Society*, 62(3):99–117, 2014.

[309] J. M. Kates and K. H. Arehart. The hearing-aid speech perception index (haspi). *Speech Communication*, 65:75–93, 2014.

[310] J. Ma, Y. Hu, and P. C. Loizou. Objective measures for predicting speech intelligibility in noisy conditions based on new band-importance functions. *Journal of the Acoustical Society of America*, 125(5):3387–3405, 2009.

[311] T. H. Falk, V. Parsa, J. F. Santos, K. Arehart, O. Hazrati, R. Huber, J. M. Kates, and S. Scollie. Objective quality and intelligibility prediction for users of assistive listening devices: Advantages and limitations of existing tools. *IEEE signal processing magazine*, 32(2):114–124, 2015.

[312] J. Kates. An auditory model for intelligibility and quality predictions. In *Proceedings of Meetings on Acoustics ICA2013*, volume 19, page 050184. ASA, 2013.

[313] Y. Tang, M. Cooke, B. M. Fazenda, and T. J. Cox. A metric for predicting binaural speech intelligibility in stationary noise and competing speech maskers a. *Journal of the Acoustical Society of America*, 140(3):1858–1870, 2016.

[314] R. Huber, A. Pusch, N. Moritz, J. Rennies, H. Schepker, and B. T. Meyer. Objective assessment of a speech enhancement scheme with an automatic speech recognition-based system. In *Proc. Speech Communication; 13th ITG-Symposium*, pages 1–5. VDE, 2018.

[315] M. Torcoli and C. Uhle. On the effect of artificial distortions on objective performance measures for dialog enhancement. In *Proc. 141st Audio Engingeering Society Convention*. Audio Engineering Society, 2016.

[316] D Sanchez. Reading speed, Aug. 2014. URL http://www.ericsson.com/broadcastandmedia/access-services/reading-speed/?platform=hootsuite.

[317] BBC. BBC Sound Effects Library CDs 1-60.

[318] Tera Media and CRG. Soundsnap.com. URL http://www.soundsnap.com/.

[319] R.C. Bilger, J.M. Nuetzel, W.M. Rabinowitz, and C. Rzeczkowski. Standardization of a test of speech perception in noise. *Journal of Speech, Language and Hearing Research*, 27(1):32–48, 1984.

[320] M. Torcoli and S. Dick. Comparing the effect of audio coding artifacts on objective quality measures and on subjective ratings. In *Proc. 144th Audio Engingeering Society Convention*. Audio Engineering Society, 2018.

[321] Gaston Hilkhuysen and Mark Huckvale. Can physical metrics identify noise reduction settings that optimize intelligibility? In *Proceedings of Meetings on Acoustics ICA2013*, volume 19, page 060119. Acoustical Society of America, 2013.

[322] Michael A Stone and Shanelle Canavan. The near non-existence of "pure" energetic masking release for speech: Extension to spectro-temporal modulation and glimpsing. *The Journal of the Acoustical Society of America*, 140(2):832–842, 2016.

[323] ITU Recommendation. ITU-R BS. 1770-2, Algorithms to measure audio programme loudness and true-peak audio level. 2011.

[324] ITU Recommendation. ITU-R BS.1116-1, Methods for the subjective assessment of small impairments in audio systems including multichannel sound systems. 1997.

[325] U. Halekoh, S. Højsgaard, J. Yan, et al. The r package geepack for generalized estimating equations. *Journal of Statistical Software*, 15(2):1–11, 2006.

[326] K. Bartoń. Mumin: multi-model inference. r package version 1.10. 0. 2014. URL http://CRAN.R-project.org/package=MuMIn.

[327] H. Zwicker. Psychoacoustics, 1990.

[328] J. B. Fritz, M. Elhilali, S. V. David, and S. A. Shamma. Auditory attention—focusing the searchlight on sound. *Current Opinion in Neurobiology*, 17(4):437–455, 2007.

[329] Joshua GW Bernstein and Ken W Grant. Auditory and auditory-visual intelligibility of speech in fluctuating maskers for normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of America*, 125(5):3358–3372, 2009.

[330] Joshua GW Bernstein and Douglas S Brungart. Effects of spectral smearing and temporal fine-structure distortion on the fluctuating-masker benefit for speech at a fixed signal-to-noise ratio. *The Journal of the Acoustical Society of America*, 130(1): 473–488, 2011.

[331] M. C. Killion, P. A. Niquette, G. I. Gudmundsen, L. J. Revit, and S. Banerjee. Development of a quick speech-in-noise test for measuring signal-to-noise ratio loss in normal-hearing and hearing-impaired listeners. *Journal of the Acoustical Society of America*, 116(4):2395–2405, 2004.

[332] S. Sharma, R. Tripathy, and U. Saxena. Critical appraisal of speech in noise tests: a systematic review and survey. *International Journal of Research in Medical Sciences*, 5(1):13, 2017.

[333] J. Swaminathan, C. R. Mason, T. M. Streeter, V. Best, E. Roverud, and G. Kidd. Role of binaural temporal fine structure and envelope cues in cocktail-party listening. *Journal of Neuroscience*, 36(31):8250–8257, 2016.

[334] I. J. Moon and S. H. Hong. What is temporal fine structure and why is it important? *Korean Journal of Audiology*, 18(1):1, 2014.

[335] O. Strelcyk and T. Dau. Relations between frequency selectivity, temporal fine-structure processing, and speech reception in impaired hearing. *Jour. Acoust. Soc. Am.*, 125(5):3328–3345, 2009.

[336] C. Füllgrabe and B. C. J. Moore. Evaluation of a method for determining binaural sensitivity to temporal fine structure (tfs-af test) for older listeners with normal and impaired low-frequency hearing. *Trends in hearing*, 21:2331216517737230, 2017.

[337] C. Füllgrabe, A. J. Harland, A. P. Sęk, and B. C. J. Moore. Development of a method for determining binaural sensitivity to temporal fine structure. *International Journal of Audiology*, 56(12):926–935, 2017.

[338] K. Hopkins and B. C. J. Moore. Development of a fast method for measuring sensitivity to temporal fine structure information at low frequencies. *International Journal of Audiology*, 49(12):940–946, 2010.

[339] B. C. J. Moore and A. Sek. Development of a fast method for determining sensitivity to temporal fine structure. *International Journal of Audiology*, 48(4):161–171, 2009.

[340] E. L. Thorndike and I. Lorge. The teacher's word book of 30,000 words. 1952.

[341] Adobe. Audition. URL https://www.adobe.com/uk/products/audition.html.

[342] Lauren Ward, Ben Shirley, Yan Tang, and William Davies. The effect of situation-specific acoustic cues on speech intelligibility in noise. In *Proc. 18th Annual Conference of International Speech Communication Association*, pages 2958–2962, Stockholm, Sweden, Aug. 2017. ISCA.

[343] Robert C. Bilger. Manual for the clinical use of the revised SPIN test.

[344] J. Barker and M. Cooke. Modelling speaker intelligibility in noise. *Speech Communication*, 49(5):402–417, 2007.

[345] Y. Tang and M. Cooke. Subjective and objective evaluation of speech intelligibility enhancement under constant energy and duration constraints. In *Proc. 12th Annual Conference of International Speech Communication Association*, pages 345–348, Florence, Italy, 2011. ISCA.

[346] T. W. Tillman and W. O. Olsen. Speech audiometry. *Modern developments in audiology*, 2:37–74, 1973.

[347] Etymotic Research Inc. QuickSIN: Speech-in-Noise Test. 1.3, 2006.

[348] P. W. Dawson, A. A. Hersbach, and B. A. Swanson. An adaptive Australian sentence test in noise (AuSTIN). *Ear and Hearing*, 34(5):592–600, 2013.

[349] W. Hu, B. A. Swanson, and G. Z. Heller. A statistical method for the analysis of speech intelligibility tests. *PloS one*, 10(7):e0132409, 2015.

[350] M. A. Lawrence. ez: Easy analysis and visualization of factorial experiments. 2012. URL https://cran.r-project.org/web/packages/ez/index.html.

[351] Cynthia A Hogan and Christopher W Turner. High-frequency audibility: Benefits for hearing-impaired listeners. *Journal of the Acoustical Society of America*, 104(1): 432–441, 1998.

[352] P. T. Johannesen, P. Pérez-González, S. Kalluri, J. L. Blanco, and E. A. Lopez-Poveda. The influence of cochlear mechanical dysfunction, temporal processing deficits, and age on the intelligibility of audible speech in noise for hearing-impaired listeners. *Trends in hearing*, 20:2331216516641055, 2016.

[353] V. Summers, M. J. Makashay, S. M. Theodoroff, and M. R. Leek. Suprathreshold auditory processing and speech perception in noise: hearing-impaired and normal-hearing listeners. *Journal of the American Academy of Audiology*, 24(4):274–292, 2013.

[354] H. Meretoja. Narrative and human existence: Ontology, epistemology, and ethics. *New Literary History*, 45(1):89–109, 2014.

[355] A. J. Cohen. the interpretation of film and video: Approaches from experimental psychology. *Selected Reports in Ethnomusicology*, 12:15, 2005.

[356] S. Deutsch. The soundtrack: Editorial. *The Soundtrack*, 1(3):183–191, 2008.

[357] J. G. Butler. *Television: Critical methods and applications*. Routledge, 2012.

[358] A. J. Cohen. The functions of music in multimedia: A cognitive approach. In *Proc. 5th International Conference on Music Perception and Cognition*, Seoul, Korea, Aug. 1998.

[359] R. S. Pellegrini and A. Krueger. Object-audio capture system for sports broadcasting. *SMPTE Motion Imaging Journal*, 128(5):46–50, 2019.

[360] R. Oldfield and B. G. Shirley. Enhanced next generation audio for live sports broadcast. In *Proc. International Broadcast Convention Conference*, Sept. 2017.

[361] H. Zettl. The rare case of television aesthetics. *Journal of the University film association*, 30(2):3–8, 1978.

[362] L. Fryer. *Putting it into words: The impact of visual impairment on perception, experience and presence*. PhD thesis, Goldsmiths, University of London, 2013.

[363] J. Whitehead. What is audio description. In *International Congress Series*, volume 1282, pages 960–963. Elsevier, 2005.

[364] S. Rai, J. Greening, and L. Petré. A comparative study of audio description guidelines prevalent in different countries. *London: Media and Culture Department, Royal National Institute of Blind People (RNIB)*, 2010.

[365] J. J. Gibson. *The ecological approach to visual perception: classic edition*. Psychology Press, 1979.

[366] T. McAlpin and D. Phillips. Jake and Sophia, 2014.

[367] William W Gaver. What in the world do we hear?: An ecological approach to auditory event perception. *Ecological psychology*, 5(1):1–29, 1993.

[368] W. W. Gaver. How do we hear in the world? explorations in ecological acoustics. *Ecological psychology*, 5(4):285–313, 1993.

[369] C. Carello, K. L. Anderson, and A. J. Kunkler-Peck. Perception of object length by sound. *Psychological science*, 9(3):211–214, 1998.

[370] S. Lakatos, S. McAdams, and R. Caussé. The representation of auditory source characteristics: Simple geometric form. *Perception & psychophysics*, 59(8):1180–1190, 1997.

[371] C. J. Steenson, M. Rodger, and W. Matthew. Bringing sounds into use: thinking of sounds as materials and a sketch of auditory affordances. *The Open Psychology Journal*, 8(1), 2015.

[372] W. L. Windsor and C. De Bézenac. Music and affordances. *Musicae scientiae*, 16(1): 102–120, 2012.

[373] E. F. Clarke et al. *Ways of listening: An ecological approach to the perception of musical meaning*. OUP USA, 2005.

[374] J. Krueger. Affordances and the musically extended mind. *Frontiers in psychology*, 4: 1003, 2014.

[375] A. Sedda, S. Monaco, G. Bottini, and M. A. Goodale. Integration of visual and auditory information for hand actions: preliminary evidence for the contribution of natural sounds to grasping. *Experimental brain research*, 209(3):365–374, 2011.

[376] D. Purser. Comparisons of evacuation efficiency and pre-travel activity times in response to a sounder and two different voice alarm messages. In *Pedestrian and Evacuation Dynamics 2008*, pages 121–134. Springer, 2010.

[377] J. Woodcock, J. Francombe, A. Franck, P. Coleman, et al. A framework for intelligent metadata adaptation in object-based audio. In *Proc. Audio Engineering Society Conference on Spatial Reproduction*, Tokyo, Japan, August 2018.

[378] L. Ward, B. Shirley, and J. Francombe. Accessible object-based audio using hierarchical narrative importance metadata. In *Proc. 145th Audio Engineering Society Convention*. Audio Engineering Society, 2018.

[379] ITU-R. Audio definition model. Technical Report BS.2076-0, 2015. URL https://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.2076-0-201506-I!!PDF-E.pdf.

[380] A. Franck and F. M. Fazi. Visr—a versatile open software framework for audio signal processing. In *Proc. Audio Engineering Society International on Spatial Reproduction*, Jul 2018. URL http://www.aes.org/e-lib/browse.cfm?elib=19628.

[381] V. Pulkki. Virtual sound source positioning using vector base amplitude panning. *Journal of the Audio Engineering Society*, 45(6):456–466, June 1997.

[382] J. Woodcock et al. Presenting the S3A object-based audio drama dataset. In *Audio Engineering Society Convention 140*, Paris, France, June 2016.

[383] A. Bryman. *Social research methods*. Oxford university press, 2004.

[384] Steinberg. Nuendo. URL https://new.steinberg.net/nuendo/.

[385] Cockos. REAPER: Rapid Environment for Audio Production, Engineering, and Recording. URL https://www.reaper.fm/.

[386] S3A: Future spatial audio for the home project. VISR- Versatile Interactive Scene Renderer Framework, 2019. URL https://www.s3a-spatialaudio.org/visr.

[387] T. Nixon, A. Bonney, and F. Melchior. A reference listening room for 3d audio research. In *3rd Int Conf on Spatial Audio,*, Graz, Austria, Sept. 2015.

[388] E. Chourdakis, L. Ward, M. Paradis, and J. Reiss. Modelling experts' decisions on assigning narrative importances of objects in a radio drama mix. In *Proc. 22nd International Conference on Digital Audio Effects - DAFX*, Sept. 2019.

[389] BBC One. Casualty. URL https://www.bbc.co.uk/programmes/b006m8wd.

[390] C. Bartlett. Standard media player, 2014. URL https://www.bbc.co.uk/blogs/internet/entries/7185ad76-d3de-3df6-8641-975feed88091.

[391] BBC. BBC iPlayer, 2019. URL https://www.bbc.co.uk/iplayer.

[392] P. Adenot and H. Choi. Web audio api, 2019. https://webaudio.github.io/web-audio-api/.

[393] B. G. Shirley, L. Ward, and E. Chourdakis. Personalization of object-based audio for accessibility using narrative importance. In *DataTV: 1st International Workshop on Data-Driven Personalisation of Television*, Manchester, UK, June 2019.

[394] Avid. Pro Tools. URL https://www.avid.com/pro-tools.

[395] NUGEN Audio. NUGEN Post. URL https://nugenaudio.com/nugenpost/.

[396] M. Moore. Can't hear mumbling dialogue on bbc? a solution is on the way. Jul. 2019. URL https://www.thetimes.co.uk/article/can-t-hear-mumbling-dialogue-on-bbc-a-solution-is-on-the-way-6kmh0zqwl.

[397] H. Goodwin and V. Bell. Tv viewers could soon finally be free of mumbling actors as bbc trials technology that turns down background noise and boosts voices. Jul. 2019. URL https://www.dailymail.co.uk/sciencetech/article-7289139/BBC-trials-technology-make-dialogue-easier-follow.html.

[398] The Radio Times. Reader report: The beginning of the end for mumbling? 10th Aug. 2019.

[399] BBC. BBC Points of View: Episode 11, Sept. 2019. URL m00085vq.

[400] G. R. Warnes, B. Bolker, T. Lumley, and R. Johnson. gmodels: Various r programming tools for model fitting. 2018. URL https://cran.r-project.org/web/packages/gmodels/index.html.

[401] J. Francome and K. Hentschel. Evaluation of an immersive audio experience using questionnaire and interaction data. In *Proc. 23rd International Congress on Acoustics*, Aachen, Germany, Sept. 2019.

[402] J. Plunkett. Heard this before? BBC chief speaks out over Happy Valley mumbling, Apr. 2016. URL https://www.theguardian.com/media/2016/apr/08/bbc-happy-valley-mumbling-jamaica-inn-sarah-lancashire.

[403] Spiegel Online. Tonprobleme beim "polizeiruf" dramaturgie top, sound flop, July 2013. URL https://www.spiegel.de/kultur/tv/ton-bei-polizeiruf-war-aus-dramaturgischen-gruenden-schlecht-a-911204.html.

[404] H. Fullerton. BBC drama SS-GB criticised for “mumbling" and bad sound quality in first episode, Feb. 2017. URL http://www.radiotimes.com/news/2017-02-26/bbc-drama-ss-gb-criticised-for-mumbling-and-bad-sound-quality-in-first-episode.

[405] W. Uricchio. Television’s first seventy-five years: The interpretive flexibility of a medium in transition. In *The Oxford handbook of film and media studies*. 2008.

[406] E. Albee. *The zoo story*. Spoken Arts, 1965.

Appendix A

Ethical approval

Ethical approval for the research completed in this PhD was gained from the University of Salford Ethics board under the following approval numbers CST 14/45 (06/10/2014) and ST16/117 (31/08/2016).

Work in Part I, Part II and Chapter 9 were conducted under the ethics approval ST16/117. The letter of approval is shown for reference.

The remaining research in this work was conducted under CST 14/45.

Research, Innovation and Academic Engagement Ethical Approval Panel

Research Centres Support Team
G0.3 Joule House
University of Salford
M5 4WT

T +44(0)161 295 5278

www.salford.ac.uk/

31 August 2016

Dear Lauren,

RE: ETHICS APPLICATION ST16/117– Understanding Television Sound: Using Object Based Audio (OBA) to optimise Intelligibility and Comprehension for Individual Listeners

Based on the information you provided, I am pleased to inform you that your application ST 16/117 has been approved.

If there are any changes to the project and/ or its methodology, please inform the Panel as soon as possible by contacting S&T-ResearchEthics@salford.ac.uk

Yours sincerely,

Prof Mohammed Arif
Chair of the Science & Technology Research Ethics Panel
Professor of Sustainability and Process Management,
School of Built Environment
University of Salford
Maxwell Building, The Crescent
Greater Manchester, UK M5 4WT
Phone: + 44 161 295 6829
Email: m.arif@salford.ac.uk

Appendix B

Survey instrument

This appendix contains the survey instrument used in Chapter 3. This includes the reduced version of the 'Speech, Spatial and Qualities of Hearing Scale' (SSQ12), the developed television experience survey items (TV10) and demographic items.

This survey instrument was distributed using the onlinesurveys.co.uk platform and is formatted in accordance with that platform.

Improving Television Sound

What is this study about?

At the University of Salford, we are researching strategies to improve the clarity and intelligibility of television speech and audio.

This research has three stages;

1. Gathering information about people's experience of speech on television.
2. Developing methods to improve clarity and intelligibility of speech on television.
3. Evaluating these methods with the general public.

You are invited to take part in stage 1, which involves completing this survey.

You will also have the opportunity to volunteer for stage 3.

What will I have to do?

This survey will ask about three main things. These are:

- questions about your experience of television audio and scenarios where you may have found speech on television hard to understand.
- questions about you, such as your age and how many hours of television you watch on average each day.
- questions about how your hearing performs in everyday situations.

It will take 15-20 mins to complete.

You can find out more about our research here.

Taking Part

Your participation in this study is voluntary and confidential.

All data will be anonymised and kept in accordance with UK Data Protection laws. It will only be used for research purposes.

*Results from partially completed surveys will not be stored and you can stop the survey at any point. Unless you click **Finish** on the final page of the survey, no data will be recorded. If you wish to withdraw your data after it has been submitted, contact L.Ward7@edu.salford.ac.uk.*

Consent to take part

I have read and understood the above information and agree to take part in this survey ✱ *Required*

- ○ Yes
- ○ No

To allowyou to withdrawyour data after you have taken party in this study, we ask you to submit two small pieces of information which will help identify you whilst maintaining your anonymity. You can choose not to provide this information, however then we will be unable to withdrawyour data.

Please enter the last three letters or digits of your postcode or zip code. *Optional*

More info

Please enter the last three digits of your telephone number. *Optional*

More info

About You

This section contains seven questions which are designed to find out some basic information about you.

The information from these questions will help us to understand the relationship between your experiences with broadcast audio and other personal characteristics, such as howoften you watch television.

1. What is your age?

More info

2. Where do you live? (country of residence)

More info

3. What is your native language?

More info

4. How many hours a day do you watch television, on average?

- Less than 1 hr
- 1 - 2hrs
- 3 - 4 hrs
- More than 4 hrs

5. What type of programming do you mostly watch? (you may select more than one option)

- News and Current Affairs programming
- Drama and Soaps
- Sports
- Documentary
- Comedy
- Lifestyle, Music and Food
- Films

6. Are you a musician?

- No
- Yes, I am a professional musician
- Yes, I am an amateur musician

7. Do you identify as Hard of Hearing? ✱ *Required*

More info

- Yes
- No
- Prefer not to say

8. What type of device are you completing this survey on?

- Mobile phone
- Tablet
- PC or Laptop
- Other

About Your Hearing

You have identified as having some degree of hearing loss in a previous question.

The six questions in this section are designed to help us better understand the degree and type of hearing loss you have. You are reminded that this survey is confidental and all data is made anonymous. If you feel uncomfortable answering any of the following questions, please leave them blank.

1. What degree of hearing loss do you have?

More info

- Mild
- Moderate
- Severe
- Profound
- I do not have hearing loss
- I'm not sure

2. Do you suffer from Tinnitus?

- No
- Yes

3. Do you use any assistive hearing devices?

More info

- No
- Yes, only in my right ear
- Yes, only in my left ear
- Yes, in both ears

*If you answered no to the above question, please go to the bottom of the page and click **Next***

4. If you answered yes above, what sort of assistive device do you use? (leave blank if not applicable)

5. Do you regularly wear this device? (leave blank if not applicable)

- Yes
- No

6. How many years have you had this device fitted? (leave blank if not applicable)

Your Experience of Television

This section contains fourteen questions. They aim to investigate your television watching patterns and howeasy you find speech on television to understand.

Those using mobile phones: All response scales are from 0 to 10, you may need to scroll right on your screen to see the full scale.

When answering these questions imagine yourself watching television in an otherwise quiet room. When background noise is mentioned in the following questions it refers to background noise in the television program, not in the environment you are watching television in.

1. Generally, how difficult do you find it to understand speech on television?

More info

Please don't select more than 1 answer(s) per row.

	0	1	2	3	4	5	6	7	8	9	10	
Very Difficult (0)	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	Very Easy (10)

2. Thinking about a recent drama you have watched on television.
Which of the following sounds helped you personally to follow the plot? (Select as many as you think apply)

- ☐ Dialogue
- ☐ Foreground sounds (e.g.the main character slamming a door in anger or other sounds that the characters can hear)
- ☐ Background sounds (e.g. sounds of birds in the countryside, background chatter in a pub scene)
- ☐ Music

3. Which of the following types of television content do you find the dialogue/speech easiest to understand whilst watching? (Select as many as you think apply)

- ☐ News and Current Affairs
- ☐ Drama and Soaps
- ☐ Sports
- ☐ Lifestyle, Music and Food
- ☐ Documentary
- ☐ Comedy
- ☐ Films

For the following questions, select any value from 0 to 10 depending on howeasily you are able to do what is described.

Selecting 10 means you are perfectly able to do what is described in the question. Selecting 0 means you would be unable to do or experience what is described.

4. A character is speaking but they are not on screen. How easily can you understand the speech without seeing the character's face?

Please don't select more than 1 answer(s) per row.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all (0)	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	Perfectly (10)

5. You are watching a panel show and one of the panellists is speaking whilst the studio audience laughs and cheers. How easily are you able to understand the panellist's speech?

Please don't select more than 1 answer(s) per row.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all (0)	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	Perfectly (10)

6. How often do you use subtitles?

Please don't select more than 1 answer(s) per row.

	0	1	2	3	4	5	6	7	8	9	10	
Never (0)	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	Always (10)

7. A news presenter is reporting from a quiet studio. Without using subtitles, how easily can you understand the speech?

Please don't select more than 1 answer(s) per row.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all (0)	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	Perfectly (10)

8. You are watching a scene on television which has the sound of clinking glasses, music and people talking in the background. Can you make out the different sounds?

Please don't select more than 1 answer(s) per row.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all (0)	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	Perfectly (10)

9. You are watching a nature documentary. The narrator is speaking with the constant sound of a waterfall in the background. Can you follow what the narrator is saying?

Please don't select more than 1 answer(s) per row.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all (0)	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	Perfectly (10)

10. How much effort do you require to hear what is being said in a television drama?

Please don't select more than 1 answer(s) per row.

	0	1	2	3	4	5	6	7	8	9	10	
A lot of effort (0)	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	No effort (10)

11. How often do you watch television programs which are not in your native language?

More info

Please don't select more than 1 answer(s) per row.

	0	1	2	3	4	5	6	7	8	9	10	
Never (0)	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	Always (10)

12. When sign-interpretation is available, how often do you watch sign language interpreted programming?

Please don't select more than 1 answer(s) per row.

	0	1	2	3	4	5	6	7	8	9	10	
Never (0)	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	Always (10)

The final two questions in this section allowyou to explain your personal experience with television sound. You can write as much or as little as you like.

13. What measures do you feel would make television speech easier for you to understand?

14. Are there any specific examples of times when you have found television speech hard to understand that you would like to tell us about?

Your Hearing

This section contains twelve questions which aim to understand more about howyour hearing performs in everyday situations.

Those using mobile phones: All response scales are from 0 to 10, you may need to scroll right on your screen to see the full scale.

For the following questions, select any value from 0 to 10 depending on howeasily you are able to do what is described.

Selecting 10 means you are perfectly able to do what is described in the question. Selecting 0 means you would be unable to do or experience what is described.

1. You are talking with one other person and there is a TV on in the same room. Without turning the TV down, can you follow what the person you're talking to says?

Please don't select more than 1 answer(s) per row.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all (0)	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	Perfectly (10)

2. You are listening to someone talking to you, while at the same time trying to follow the news on TV. Can you follow what both people are saying?

Please don't select more than 1 answer(s) per row.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all (0)	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	Perfectly (10)

3. You are in conversation with one person in a room where there are many other people talking. Can you follow what the person you are talking to is saying?

Please don't select more than 1 answer(s) per row.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all (0)	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	Perfectly (10)

4. You are in a group of about five people in a busy restaurant. You can see everyone else in the group. Can you follow the conversation?

Please don't select more than 1 answer(s) per row.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all (0)	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	Perfectly (10)

5. You are with a group and the conversation switches from one person to another. Can you easily follow the conversation without missing the start of what each new speaker is saying?

Please don't select more than 1 answer(s) per row.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all (0)	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	Perfectly (10)

6. You are outside. A dog barks loudly. Can you tell immediately where it is, without having to look?

Please don't select more than 1 answer(s) per row.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all (0)	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	Perfectly (10)

7. Can you tell how far away a bus or a truck is, from the sound?

Please don't select more than 1 answer(s) per row.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all (0)	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	Perfectly (10)

8. Can you tell from the sound whether a bus or truck is coming towards you or going away?

Please don't select more than 1 answer(s) per row.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all (0)	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	Perfectly (10)

9. When you hear more than one sound at a time, do you have the impression that it seems like a single jumbled sound?

Please don't select more than 1 answer(s) per row.

	0	1	2	3	4	5	6	7	8	9	10	
Jumbled (0)	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	Not Jumbled (10)

10. When you listen to music, can you make out which instruments are playing?

Please don't select more than 1 answer(s) per row.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all (0)	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	Perfectly (10)

11. Do everyday sounds that you can hear easily seem clear to you (not blurred)?

Please don't select more than 1 answer(s) per row.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all (0)	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	Perfectly (10)

12. Do you have to concentrate very much when listening to someone or something?

Please don't select more than 1 answer(s) per row.

	0	1	2	3	4	5	6	7	8	9	10	
Concentrate (0)	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	No need to concentrate (10)

Further Participation

*If you would like to be contacted with opportunities to volunteer for stage 3 (in person listening tests at the University of Salford) or simply stay up to date with our research, please answer **yes** below and input your details. Your participation helps us to develop and test new methods to improve television sound and ensure that it meets user needs.*

*Otherwise select **no** below and **Finish** to complete the survey.*

Would you like to be informed of opportunities to volunteer for future listening studies? ✱ *Required*

- ○ Yes
- ○ No

Would you like to join our mailing list to receive updates about our research? *Optional*

- ○ Yes
- ○ No

Please input your telephone number or email address, if you answered yes above.

Appendix C

Modified R-SPIN sentence lists

This appendix contains the sentence lists modified from the Revised Speech Perception in Noise (R-SPIN) test [82] which were used for the single speech to background ratio speech in noise tests in Study One, Two and Three in Chapter 5 and Chapter 6.2.

Description of this methodology and the rationale for its selection is described in Chapter 4.

The following Tables show the sentences used in each list, with the keyword for recognition noted in bold. Whether the sentence is high or low predictability and whether there was a redundant, non-speech audio object (SFX) included is then noted.

C.1 List One

	Sentence	Keyword	Predictability		SFX
1.	His plan meant taking a big risk.	RISK	H		–
2.	Stir your coffee with a spoon.	SPOON	H		SFX
3.	Miss White won't think about the crack.	CRACK		L	–
4.	He would think about the rag.	RAG		L	–
5.	The plough was pulled by an ox.	OX	H		SFX
6.	The old train was powered by steam.	STEAM	H		SFX
7.	The old man talked about the lungs.	LUNGS		L	SFX
8.	I was considering the crook.	CROOK		L	–
9.	Let's decide by tossing a coin.	COIN	H		SFX
10.	The doctor prescribed the drug.	DRUG	H		–
11.	Bill might discuss the foam.	FOAM		L	–
12.	Nancy didn't discuss the skirt.	SKIRT		L	–
13.	Hold the baby on your lap.	LAP	H		–
14.	Bob has discussed the splash.	SPLASH		L	SFX
15.	The dog chewed on a bone.	BONE	H		–
16.	Ruth hopes he heard about the hips.	HIPS		L	–
17.	The war was fought with armoured tanks.	TANKS	H		SFX
18.	She wants to talk about the crew.	CREW		L	–
19.	They had a problem with the cliff.	CLIFF		L	–
20.	They drank a whole bottle of gin.	GIN	H		SFX
21.	You heard Jane called about the van.	VAN		L	SFX
22.	The witness took a solemn oath.	OATH	H		–
23.	We could consider the feast.	FEAST		L	SFX
24.	Bill heard we asked about the host.	HOST		L	–
25.	They tracked the lion to his den.	DEN	H		–
26.	The cow gave birth to the calf.	CALF	H		SFX
27.	I hod not thought about the growl.	GROWL		L	SFX
28.	The scarf was made of silk.	SILK	H		–
29.	The super highway had six lanes.	LANES	H		SFX
30.	The old man discussed the yell.	YELL		L	SFX
31.	The shepherd watched his flock of sheep.	SHEEP	H		SFX
32.	The beer drinkers raised their mugs.	MUGS	H		SFX
33.	I'm glad you heard about the bend.	BEND		L	–
34.	You're talking about the pond.	POND		L	SFX
35.	The rude remark made her blush.	BLUSH	H		–
36.	Nancy had considered the sleeves.	SLEEVES		L	–
37.	We heard the ticking of the clock.	CLOCK	H		SFX
38.	He can't consider the crib.	CRIB		L	SFX
39.	He killed the dragon with his sword.	SWORD	H		SFX
40.	Tom discussed the hay.	HAY		L	–
41.	Mary wore her hair in braids.	BRAIDS	H		–
42.	She's glad Jane asked about the drain.	DRAIN		L	SFX
43.	Bill hopes Paul heard about the mist.	MIST		L	–
44.	Greet the heroes with loud cheers.	CHEERS	H		SFX
45.	No one was injured in the crash.	CRASH	H		SFX
46.	We're speaking about the toll.	TOLL		L	–
47.	My son has a dog as a pet.	PET	H		SFX
48.	He was scared out of his wits.	WITS	H		–
49.	Bob should consider the mice.	MICE		L	SFX
50.	I've spoken about the pile.	PILE	H		–

C.2 List Two

	Sentence	Keyword	Predictability		SFX
1.	Miss Black thought about the lap.	LAP		L	–
2.	The baby slept in his crib.	CRIB	H		SFX
3.	The watchdog gave a warning growl.	GROWL	H		SFX
4.	Miss Black would consider the bone.	BONE		L	–
5.	The drowning man let out a yell.	YELL	H		SFX
6.	Bob could have known about the spoon.	SPOON		L	SFX
7.	Our cat is good at catching mice.	MICE	H		SFX
8.	He wants to talk about the risk	RISK		L	–
9.	He heard they called about the lanes.	LANES		L	SFX
10.	Wipe your greasy hands on the rag.	RAG	H		–
11.	She has known about the drug.	DRUG		L	–
12.	I want to speak about the crash.	CRASH		L	SFX
13.	The wedding banquet was a feast.	FEAST	H		SFX
14.	We are considering the cheers.	CHEERS		L	SFX
15.	Paul hit the water with a splash.	SPLASH	H		SFX
16.	The ducks swam in the pond.	POND	H		SFX
17.	They've considered the sheep.	SHEEP		L	SFX
18.	The man should discuss the ox.	OX		L	SFX
19.	Bob stood with his hands on his hips.	HIPS	H		–
20.	The cigarette smoke filled his lungs.	LUNGS	H		SFX
21.	They heard I called her about the pet.	PET		L	SFX
22.	The cushion was filled with foam.	FOAM	H		–
23.	Ruth poured the water down the drain.	DRAIN	H		SFX
24.	Bill cannot consider the den.	DEN		L	–
25.	The nozzle sprays a fine mist.	MIST	H		–
26.	The sport shirt has short sleeves.	SLEEVES	H		–
27.	She hopes Jane called about the calf.	CALF		L	SFX
28.	Jane has a problem with the coin.	COIN		L	SFX
29.	She shortened the hem on her skirt.	SKIRT	H		–
30.	Paul hopes she calls about the tanks.	TANKS		L	SFX
31.	The girl talked about the gin.	GIN		L	SFX
32.	The guests were welcomed by the host.	HOST	H		–
33.	Mary should think about the sword.	SWORD		L	SFX
34.	Ruth could have discussed the wits.	WITS		L	–
35.	The ship's captain summoned his crew.	CREW	H		–
36.	You had a problem with a blush.	BLUSH		L	–
37.	The flood took a heavy toll.	TOLL	H		–
38.	The car drove off the steep cliff.	CLIFF	H		–
39.	We have discussed the steam.	STEAM		L	SFX
40.	The policeman captured the crook.	CROOK	H		–
41.	The door was open just a crack.	CRACK	H		–
42.	Tom is considering the clock.	CLOCK		L	SFX
43.	The sand was heaped in a pile.	PILE	H		–
44.	You should not speak about the braids.	BRAIDS		L	–
45.	Peter should speak about the mugs.	MUGS	H		SFX
46.	Households goods are moved in a van.	VAN		L	SFX
47.	He has a problem with the oath.	OATH		L	–
48.	Follow this road around the bend.	BEND	H		–
49.	Tom won't consider the silk.	SILK		L	–
50.	The farmer baled his hay.	HAY	H		–

C.3 List Three

	Sentence	Keyword	Predictability	SFX
1.	The woman talked about the frogs.	FROGS	L	SFX
2.	You'd been considering the geese.	GEESE	L	SFX
3.	It's getting dark, so light the lamp.	LAMP	H	–
4.	Kill the bugs with this spray.	SPRAY	H	SFX
5.	They heard I asked about the bet.	BET	L	–
6.	The mouse was caught in the trap.	TRAP	H	SFX
7.	Mary knows about the rug.	RUG	L	–
8.	The airplane went into a dive.	DIVE	H	SFX
9.	The fireman heard her frightened scream.	SCREAM	H	SFX
10.	He was interested in the hedge.	HEDGE	L	–
11.	He wiped the sink with a sponge.	SPONGE	H	–
12.	He is considering the throat.	THROAT	L	SFX
13.	Tom discussed the swan.	SWAN	L	SFX
14.	The papers were held by a clip.	CLIP	H	–
15.	Paul can't discuss the wax.	WAX	L	–
16.	The old man considered the kick.	KICK	L	SFX
17.	The chicks followed the mother hen.	HEN	H	SFX
18.	David might consider the fun.	FUN	L	–
19.	She wants to speak about the ant.	ANT	L	–
20.	The airplane dropped a bomb.	BOMB	H	SFX
21.	The boy took shelter in a cave.	CAVE	H	–
22.	He hasn't considered the dart.	DART	L	SFX
23.	Eve was made with Adam's rib.	RIB	H	–
24.	We shipped the furniture by truck.	TRUCK	H	SFX
25.	He is thinking about the roar.	ROAR	L	SFX
26.	The girl swept the floor with a broom.	BROOM	H	SFX
27.	Miss White thinks about the tea.	TEA	L	SFX
28.	Jane didn't think about the brook.	BROOK	L	SFX
29.	Cut a piece of meat from the roast.	ROAST	H	–
30.	Betty can't consider the grief.	GRIEF	L	SFX
31.	The heavy rains caused a flood.	FLOOD	H	–
32.	The swimmer dove into the pool.	POOL	H	SFX
33.	Harry will consider the trail.	TRAIL	L	–
34.	Let's invite the whole gang.	GANG	H	–
35.	The house was robbed by a thief.	THIEF	H	–
36.	Tom is talking about the fee.	FEE	L	–
37.	Bob wore a watch on his wrist.	WRIST	H	–
38.	Tom had spoken about the pill.	PILL	L	–
39.	Ann was interested in the breath.	BREATH	L	SFX
40.	The secret agent was a spy.	SPY	H	–
41.	The rancher rounded up his herd.	HERD	H	SFX
42.	Tom couldn't have thought about the sport.	SPORT	H	SFX
43.	Mary can't consider the tide.	TIDE	L	SFX
44.	Ann works in the bank as a clerk.	CLERK	H	–
45.	A chimpanzee is an ape.	APE	H	SFX
46.	He hopes Tom asked about the bar.	BAR	L	–
47.	Mr White discussed the cruise.	CRUISE	L	SFX
48.	The bandits escaped from jail.	JAIL	H	SFX
49.	Paul hopes we heard about the loot.	LOOT	L	–
50.	The landlord raised the rent.	RENT	H	–

C.4 List Four

	Sentence	Keyword	Predictability	SFX
1.	You were considering the gang.	GANG	L	–
2.	She's spoken about the bomb.	BOMB	L	SFX
3.	Playing checkers can be fun.	FUN	H	–
4.	The doctor charged a low fee.	FEE	H	–
5.	He wants to know about the rib.	RIB	L	–
6.	The gambler lost the bet	BET	H	–
7.	I've got a cold and a sore throat.	THROAT	H	SFX
8.	She might have discussed the ape.	APE	H	SFX
9.	The woman talked about the frogs.	FROGS	L	SFX
10.	Instead of a fence, plant a hedge.	HEDGE	H	–
11.	The old woman discussed the thief.	THIEF	L	–
12.	The duck swam with the white swan.	SWAN	H	SFX
13.	They fished in the babbling brook.	BROOK	H	SFX
14.	You were interested in the scream.	SCREAM	L	SFX
15.	Mary had considered the spray.	SPRAY	L	SFX
16.	The widow's sob expressed her grief.	GRIEF	H	SFX
17.	The candle's flame melted the wax.	WAX	H	–
18.	I haven't discussed the sponge.	SPONGE	L	–
19.	He was hit by a poisoned dart.	DART	H	SFX
20.	How long can you hold your breath.	BREATH	H	SFX
21.	Ruth will consider the herd.	HERD	L	SFX
22.	Ruth poured herself a cup of tea.	TEA	H	SFX
23.	The old man discussed the dive	DIVE	L	SFX
24.	The class should consider the flood.	FLOOD	L	–
25.	The lion gave an angry roar.	ROAR	H	SFX
26.	The girl swept the floor with a broom.	BROOM	L	SFX
27.	Paul has discussed the lamp.	LAMP	L	–
28.	We saw a flock of wild geese.	GEESE	H	SFX
29.	You knew about the clip.	CLIP	L	–
30.	She might consider the pool.	POOL	L	SFX
31.	We swan at the beach at high tide.	TIDE	H	SFX
32.	Bob was considering the clerk.	CLERK	L	–
33.	We got drunk in the local bar.	BAR	H	–
34.	A termite looks like an ant.	ANT	H	–
35.	The man knew about the spy.	SPY	L	–
36.	The sick child swallowed the pill.	PILL	H	–
37.	The call discussed the wrist.	WRIST	L	–
38.	The burglar escaped with the loot.	LOOT	H	–
39.	They hope he heard about the rent.	RENT	L	–
40.	Mr. White spoke about the jail.	JAIL	L	SFX
41.	The steamship left on a cruise.	CRUISE	H	SFX
42.	We've spoke about the truck.	TRUCK	L	SFX
43.	Bill didn't discuss the hen.	HEN	L	SFX
44.	The bloodhound followed the trail.	TRAIL	H	–
45.	The boy considered the trap.	TRAP	L	SFX
46.	The boy gave the football a kick.	KICK	H	SFX
47.	He should consider the roast.	ROAST	L	–
48.	Miss Brown spoke about the cave.	CAVE	L	–
49.	She hated to vacuum the rug.	RUG	H	–
50.	Football is a dangerous sport.	SPORT	H	–

Appendix D

Modified R-SPIN Sentence Lists - Multiple SBR Paradigm

This appendix contains the sentence lists modified from the Revised Speech Perception in Noise (R-SPIN) test [82] which were used for the multiple speech to background ratio speech in noise tests in Study Four in Chapter 6.2.

Description of this methodology, its design and the rationale for its selection is described in Chapter 7.

The following Tables first show the practise list followed by the three main lists developed. For each list the sentence is shown, with the keyword for recognition noted in bold. Whether the sentence is high or low predictability and whether there was a redundant, non-speech audio object (SFX) included is then noted. Finally the speech to background ratio (SBR) set for each sentence is given.

D.1 Practice List

	Sentence	Keyword	Predictability	SFX	SBR
1	I cut my finger with a knife.	KNIFE	H	SFX	29
2	The fireman heard her frightened scream.	SCREAM	H	SFX	29
3	The dealer shuffled the cards.	CARDS	H	SFX	29
4	The mouse was caught in a trap.	TRAP	H	SFX	29
5	The bandits escaped from jail.	JAIL	H	SFX	29
6	They fished in the babbling brook.	BROOK	H	SFX	29
7	Use this spray to kill the bugs.	BUGS	H	SFX	23
8	Get the bread and cut me a slice.	SLICE	H	SFX	23
9	Heavy rains caused a flood.	FLOOD	H	SFX	23
10	Ruth had a necklace of glass beads.	BEADS	H	SFX	23
11	Our cat is good at catching mice.	MICE	H	SFX	23
12	The swimmer dove into the pool.	POOL	H	SFX	23
13	The rancher rounded up his herd.	HERD	H	SFX	17
14	We swam at the beach at high tide.	TIDE	H	SFX	17
15	The airplane dropped a bomb.	BOMB	H	SFX	17
16	Kill the bugs with this spray.	SPRAY	H	SFX	17
17	The boat sailed along the coast.	COAST	H	SFX	17
18	The airplane went into a dive.	DIVE	H	SFX	17
19	The lion gave an angry roar.	ROAR	H	SFX	11
20	That accident gave me a scare.	SCARE	H	SFX	11
21	He was hit by a poisoned dart.	DART	H	SFX	11
22	The chicks followed the mother hen.	HEN	H	SFX	11
23	The doctor X-rayed his chest.	CHEST	H	SFX	11
24	The widow's sob expressed her grief.	GRIEF	H	SFX	11
25	The sand was heaped in a pile.	PILE	H	SFX	5
26	A chimpanzee is an ape.	APE	H	SFX	5
27	He hit me with a clenched fist.	FIST	H	SFX	5
28	I've got a cold and a sore throat.	THROAT	H	SFX	5
29	We saw a flock of wild geese.	GEESE	H	SFX	5
30	To open the jar, twist the lid.	LID	H	SFX	5

D.2 List One

	Sentence	Keyword	Predictability		SFX	SBR
1	We're lost so let's look at the map.	MAP	H		–	14
2	Let's decide by tossing a coin.	COIN	H		SFX	14
3	We're considering the brow	BROW		L	–	14
4	Household goods are moved in a van.	VAN	H		SFX	14
5	The scarf was made from shiny silk.	SILK	H		–	14
6	Jane was interested in the stamp.	STAMP		L	–	14
7	The cushion was filled with foam.	FOAM	H		–	11
8	Greet the heroes with loud cheers.	CHEERS	H		SFX	11
9	Throw out all this useless junk.	JUNK	H		–	11
10	I've been considering the crown.	CROWN		L	–	11
11	The cow gave birth to a calf.	CALF	H		SFX	11
12	I should have known about the gum.	GUM		L	–	11
13	He wants to talk about the risk.	RISK		L	–	8
14	The marksman took careful aim.	AIM	H		–	8
15	We're speaking about the toll.	TOLL		L	–	8
16	The bottle was sealed with a cork.	CORK	H		SFX	8
17	The detective searched for a clue.	CLUE	H		–	8
18	The drowning man let out a yell.	YELL	H		SFX	8
19	The beer drinkers raised their mugs.	MUGS	H		SFX	5
20	She shortened the hem on her skirt.	SKIRT	H		–	5
21	The witness took a solemn oath.	OATH	H		–	5
22	The steamship left on a cruise.	CRUISE	H		SFX	5
23	Miss Smith knows about the tub.	TUB		L	–	5
24	Ruth must have known about the pie.	PIE		L	–	5
25	The crook entered a guilty plea.	PLEA	H		–	2
26	The watchdog gave a warning growl.	GROWL	H		SFX	2
27	She has known about the drug.	DRUG		L	–	2
28	The ship's Captain summoned his crew.	CREW	H		–	2
29	The sleepy child took a nap.	NAP	H		SFX	2
30	I've spoken about the pile.	PILE		L	–	2
31	Mary wore her hair in braids.	BRAIDS	H		–	-1
32	The war was fought with armored tanks.	TANKS	H		SFX	-1
33	Miss White won't think about the crack.	CRACK		L	–	-1
34	The ducks swan around on the pond.	POND	H		SFX	-1
35	Tom discussed the hay.	HAY		L	–	-1
36	Follow this road around the bend.	BEND	H		–	-1

D.3 List Two

	Sentence	Keyword	Predictability	SFX	SBR
1	The wedding banquet was a feast.	FEAST	H	SFX	14
2	That accident gave me a scare.	SCARE	H	–	14
3	We heard the ticking clock.	CLOCK	H	SFX	14
4	David wiped the sweat from his brow.	BROW	H	–	14
5	Tom won't consider the silk.	SILK	L	–	14
6	Mr Smith spoke about the aid.	AID	L	–	14
7	The cigarette smoke filled his lungs.	LUNGS	H	SFX	11
8	Miss Black would consider the bone.	BONE	L	–	11
9	The cow gave birth to a calf.	CALF	H	SFX	11
10	Peter knows about the raft.	RAFT	L	–	11
11	The king wore a golden crown.	CROWN	H	–	11
12	My jaw aches when i chew gum.	GUM	H	–	11
13	His plan meant taking a big risk.	RISK	H	–	8
14	The shepherd watched his flock of sheep.	SHEEP	H	SFX	8
15	The flood took a heavy toll.	TOLL	H	–	8
16	Nancy had considered the sleeves.	SLEEVES	L	–	8
17	My son has a dog for a pet.	PET	H	SFX	8
18	Bill heard we asked about the host.	HOST	L	–	8
19	Hold the baby on your lap.	LAP	H	–	5
20	Stir your coffee with a spoon.	SPOON	H	SFX	5
21	He has a problem with the oath.	OATH	L	–	5
22	Bill cannot consider the den.	DEN	L	–	5
23	Paul took a bath in the tub.	TUB	H	–	5
24	The pond was full of croaking frogs.	FROGS	H	SFX	5
25	I can't consider the plea.	PLEA	L	–	2
26	The girl swept the floor with a broom.	BROOM	H	SFX	2
27	Mr. Black considered the fleet.	FLEET	L	–	2
28	We shipped the furniture by truck.	TRUCK	H	SFX	2
29	The sand was heaped in a pile.	PILE	H	–	2
30	The cookies were kept in a jar.	JAR	H	–	2
31	The duck swam with the white swan.	SWAN	H	SFX	-1
32	We spoke about the knob.	KNOB	L	–	-1
33	The door was opened just a crack.	CRACK	H	–	-1
34	The polivemen captured the crook.	CROOK	H	–	-1
35	Ruth poured herself a cup of tea.	TEA	H	SFX	-1
36	I'm glad you heard about the bend.	BEND	L	–	-1

D.4 List Three

	Sentence	Keyword	Predictability	SFX	SBR
1	I should have considered the map.	MAP	L	–	14
2	Airmail requires a special stamp.	STAMP	H	–	14
3	Miss Smith considered the scare.	SCARE	L	–	14
4	The nurse gave him first aid.	AID	H	–	14
5	How long can you hold your breath?	BREATH	H	SFX	14
6	The railroad train ran off the track.	TRACK	H	SFX	14
7	Bill might discuss the foam.	FOAM	L	–	11
8	The dog chewed on a bone.	BONE	H	–	11
9	Ruth hopes she called about the junk.	JUNK	L	–	11
10	The old train was powered by steam.	STEAM	H	SFX	11
11	The shipwrecked sailors built a raft.	RAFT	H	–	11
12	No one was injured in the crash.	CRASH	H	SFX	11
13	Ruth poured the water down the drain.	DRAIN	H	SFX	8
14	They want to know about the aim.	AIM	L	–	8
15	The bottle was sealed with a cork.	CORK	H	SFX	8
16	The sport shirt had short sleeves.	SLEEVES	H	–	8
17	The main spoke about the clue.	CLUE	L	–	8
18	The guests were welcomed by the host.	HOST	H	–	8
19	Miss Black thought about the lap.	LAP	L	–	5
20	Nancy didn't discuss the skirt.	SKIRT	L	–	5
21	Bob was cut by the jackknife's blade.	BLADE	H	SFX	5
22	They tracked the lion to his den.	DEN	H	–	5
23	The boy gave the football a kick.	KICK	H	SFX	5
24	For desert he had apple pie.	PIE	H	–	5
25	The plow was pulled by an ox.	OX	H	SFX	2
26	The doctor prescribed the drug.	DRUG	H	–	2
27	The Admiral commands the fleet.	FLEET	H	–	2
28	She wants to talk about the crew.	CREW	L	–	2
29	He killed the dragon with his sword.	SWORD	H	SFX	2
30	He's glad you called about the jar.	JAR	L	–	2
31	You should not speak about the braids.	BRAIDS	L	–	-1
32	I was considering the crook.	CROOK	L	–	-1
33	Paul hit the water with a splash.	SPLASH	H	SFX	-1
34	Unlock the door and turn the knob.	KNOB	H	–	-1
35	The farmer baled his hay.	HAY	H	–	-1
36	They drank a whole bottle of gin.	GIN	H	SFX	-1

Appendix E

Speech, Spatial and Qualities of Hearing Scale

This appendix contains a reproduction of the full list of questions from the Speech, Spatial and Qualities of Hearing Scale by Gatehouse and Noble [249]. This survey is used in it's entirety in Chapter 8 and in it's reduced form in Chapter 3. The questions contained in the reduced form can be seen in Appendix B, along with the television experience and demographic questions used in the survey in Chapters 3 and 8.

Appendix 1. A sample item from the SSQ Questionnaire

1. You are talking with one other person and there is a TV on in the same room. Without turning the TV down, can you follow what the person you're talking to says?	Not at all — Perfectly  Min — Max	tick if not applicable ☐ ☐ or wouldn't hear it

Appendix 2. Summary of the SSQ Items

SSQ Item	*Vignette*	*Anchors*
Speech 1	You are talking with one other person and there is a TV on in the same room. Without turning the TV down, can you follow what the person you're talking to says?	Not at all–Perfectly
Speech 2	You are talking with one other person in a quiet, carpeted lounge-room. Can you follow what the other person says?	Not at all–Perfectly
Speech 3	You are in a group of about five people, sitting round a table. It is an otherwise quiet place. You can see everyone else in the group. Can you follow the conversation?	Not at all–Perfectly
Speech 4	You are in a group of about five people in a busy restaurant. You can see everyone else in the group. Can you follow the conversation?	Not at all–Perfectly
Speech 5	You are talking with one other person. There is continuous background noise, such as a fan or running water. Can you follow what the person says?	Not at all–Perfectly
Speech 6	You are in a group of about five people in a busy restaurant. You *cannot* see everyone else in the group. Can you follow the conversation?	Not at all–Perfectly
Speech 7	You are talking to someone in a place where there are a lot of echoes, such as a church or railway terminus building. Can you follow what the other person says?	Not at all–Perfectly
Speech 8	Can you have a conversation with someone whose voice is the same pitch as that of the person you're talking with?	Not at all–Perfectly
Speech 9	Can you have a conversation with someone whose voice is a different pitch from that of the person you're talking with?	Not at all–Perfectly
Speech 10	You are listening to someone talking to you, while at the same time trying to follow the news on TV. Can you follow what both people are saying?	Not at all–Perfectly
Speech 11	You are in conversation with one person in a room where there are many other people talking. Can you follow what the person you are talking to is saying?	Not at all–Perfectly
Speech 12	You are with a group and the conversation switches from one person to another. Can you easily follow the conversation without missing the start of what each new speaker is saying?	Not at all–Perfectly
Speech 13	Can you easily have a conversation on the telephone?	Not at all–Perfectly
Speech 14	You are listening to someone on the telephone and someone next to you starts talking. Can you follow what's being said by both speakers?	Not at all–Perfectly

SSQ Item	*Vignette*	*Anchors*
Spatial 1	You are outdoors in an unfamiliar place. You hear someone using a lawnmower. You can't see where they are. Can you tell right away where the sound is coming from?	Not at all–Perfectly
Spatial 2	You are sitting around a table or at a meeting with several people. You can't see everyone. Can you tell where any person is as soon as they start speaking?	Not at all–Perfectly
Spatial 3	You are sitting in between two people. One of them starts to speak. Can you tell right away whether it is the person on your left or your right, without having to look?	Not at all–Perfectly
Spatial 4	You are in an unfamiliar house. It is quiet. You hear a door slam. Can you tell right away where that sound came from?	Not at all–Perfectly
Spatial 5	You are in the stairwell of a building with floors above and below you. You can hear sounds from another floor. Can you readily tell where the sound is coming from?	Not at all–Perfectly
Spatial 6	You are outside. A dog barks loudly. Can you tell immediately where it is, without having to look?	Not at all–Perfectly
Spatial 7	You are standing on the footpath of a busy street. Can you hear right away which direction a bus or truck is coming from before you see it?	Not at all–Perfectly
Spatial 8	In the street, can you tell how far away someone is, from the sound of their voice or footsteps?	Not at all–Perfectly
Spatial 9	Can you tell how far away a bus or truck is, from the sound?	Not at all–Perfectly
Spatial 10	Can you tell from the sound which direction a bus or truck is moving, e.g. from your left to your right or right to left?	Not at all–Perfectly
Spatial 11	Can you tell from the sound of their voice or footsteps which direction a person is moving, e.g. from your left to your right or right to left?	Not at all–Perfectly
Spatial 12	Can you tell from their voice or footsteps whether the person is coming towards you or going away?	Not at all–Perfectly
Spatial 13	Can you tell from the sound whether a bus or truck is coming towards you or going away?	Not at all–Perfectly
Spatial 14	Do the sounds of things you are able to hear seem to be inside your head rather than out there in the world?	Inside my head–Out there
Spatial 15	Do the sounds of people or things you hear, but cannot see at first, turn out to be closer than expected when you do see them?	Much closer–Not closer
Spatial 16	Do the sounds of people or things you hear, but cannot see at first, turn out to be further away than expected when you do see them?	Much further–Not further
Spatial 17	Do you have the impression of sounds being exactly where you would expect them to be?	Not at all–Perfectly
Qualities 1	Think of when you hear two things at once, e.g. water running into a basin (a power tool being used) (a plane flying past) and, at the same time, a radio playing (the sound of hammering) (a truck driving past). Do you have the impression of these as sounding separate from each other?	Not at all–Perfectly
Qualities 2	When you hear more than one sound at a time, do you have the impression that it seems like a single jumbled sound?	Jumbled–Not jumbled
Qualities 3	You are in a room and there is music on the radio. Someone else in the room is talking. Can you hear the voice as something separate from the music?	Not at all–Perfectly

SSQ Item	*Vignette*	*Anchors*
Qualities 4	Do you find it easy to recognize different people you know by the sound of each one's voice?	Not at all–Perfectly
Qualities 5	Do you find it easy to distinguish different pieces of music that you are familiar with?	Not at all–Perfectly
Qualities 6	Can you tell the difference between different sounds, e.g. a car versus a bus, or water boiling in a pot versus food cooking in a frying plan?	Not at all–Perfectly
Qualities 7	When you listen to music, can you make out which instruments are playing?	Not at all–Perfectly
Qualities 8	When you listen to music, does it sound clear and natural?	Not at all–Perfectly
Qualities 9	Do everyday sounds that you can hear easily seem clear to you (not blurred)?	Not at all–Perfectly
Qualities 10	Do other people's voices sound clear and natural?	Not at all–Perfectly
Qualities 11	Do everyday sounds that you hear seem to have an artificial or unnatural quality?	Unnatural–Natural
Qualities 12	Does your own voice sound natural to you?	Not at all–Perfectly
Qualities 13	Can you easily judge another person's mood from the sound of their voice?	Not at all–Perfectly
Qualities 14	Do you have to concentrate very much when listening to someone or something?	Concentrate Hard–Not need to concentrate
Qualities 15	If you turn one hearing aid/implant off, and do not adjust the other, does everything sound unnaturally quiet? (not relevant for unaided condition)	Too quiet–Not too quiet
Qualities 16	When you are the driver in a car, can you easily hear what someone is saying who is sitting alongside you?	Not at all–Perfectly
Qualities 17	When you are a passenger, can you easily hear what the driver is saying when sitting alongside you?	Not at all–Perfectly
Qualities 18	Do you have to put in a lot of effort to hear what is being said in conversation with others?	Lot of effort–No effort
Qualities 19	Can you easily ignore other sounds when trying to listen to something?	Not easily ignore–Easily ignore

Appendix F

Accessibility in music

"While I'm here can I raise another point? That's not directly connected.

Music. Now I have enormous difficulty with anything after the Beatles, in understanding any of the lyrics at all. The way people mix things and you know, I hate to live in the past. 80 years ago, people managed to get a lot of this right. They'd invented a thing called a dialogue equalizer and that was back in the 1929, 1930. But it's so simple, you drop the bass and you boost the middle and you make sure there aren't too many people clomping around the studio and it worked. And the point is that I can still watch and understand 'An evening of comedy', I could still watch and understand 'The Third Man', 'Captain Black' which I've got on tape at home. I can watch that with pleasure and that stuff is eighty years old. And it still works.

But in music, now I'm missing out on an entire batch of contemporary culture. The leading people of the day are these hip hop and rap artists and when you look at their work on the page, you have to admit, it displays a great deal of ingenuity and insight. But when they do it in performance, it's a meaningless gabble to me. Largely because of speed, their enunciation isn't too bad. But, the rapidity and the use of idioms and, of course, the very intrusive rhythm backing that they have, drives it all away. Now, I don't think I'm going to go to their concerts and I don't think I'm going to join their fan clubs but at my age, but I don't want to miss out completely. Just because I've reached the age of 95 I don't want to turn myself off."

– Focus group participant, 14/07/2018