

# Concurrency in auditory displays for connected television

Alistair Francis Hinde

Doctor of Philosophy

UNIVERSITY OF YORK  
ELECTRONICS

September 2016



# Abstract

Many television experiences depend on users being both willing and able to visually attend to screen-based information. Auditory displays offer an alternative method for presenting this information and could benefit all users. This thesis explores how this may be achieved through the design and evaluation of auditory displays involving varying degrees of concurrency for two television use cases: menu navigation and presenting related content alongside a television show.

The first study, on the navigation of auditory menus, looked at onset asynchrony and word length in the presentation of spoken menus. The effects of these on task duration, accuracy and workload were considered. Onset asynchrony and word length both caused significant effects on task duration and accuracy, while workload was only affected by onset asynchrony. An optimum asynchrony was identified, which was the same for both long and short words, but better performance was obtained with the shorter words that no longer overlapped.

The second experiment investigated how disruption, workload, and preference are affected when presenting additional content accompanying a television programme. The content took the form of sound from different spatial locations or as text on a smartphone and the programme's soundtrack was either modified or left unaltered. Leaving the soundtrack unaltered or muting it negatively impacted user experience. Removing the speech from the television programme and presenting the secondary content as sound from a smartphone was the best auditory approach. This was found to compare well with the textual presentation, resulting in less visual disruption and imposing a similar workload.

Additionally, the thesis reviews the state-of-the-art in television experiences and auditory displays. The human auditory system is introduced and important factors in the concurrent presentation of speech are highlighted. Conclusions about the utility of concurrency within auditory displays for television are made and areas for further work are identified.



# Table of contents

<b>Abstract</b>	<b>3</b>
<b>Table of Contents</b>	<b>5</b>
<b>List of Figures</b>	<b>11</b>
<b>List of Tables</b>	<b>15</b>
<b>Acknowledgements</b>	<b>17</b>
<b>Declaration</b>	<b>19</b>
<b>1 Introduction</b>	<b>21</b>
1.1 Overview: The dominance of vision in television displays . . . . .	21
1.2 Users effected by visio-centric design . . . . .	23
1.3 Auditory display . . . . .	25
1.4 Hypothesis . . . . .	25
1.5 Contributions . . . . .	26
1.6 Structure of thesis . . . . .	27
<b>2 Television experiences: Now and the near future</b>	<b>29</b>
2.1 Introduction . . . . .	29
2.2 Programme delivery . . . . .	30
2.3 Television devices . . . . .	31
2.4 Elements of the television experience . . . . .	33
2.4.1 Navigation and choice . . . . .	34
2.4.2 Additional media activities . . . . .	38
2.4.3 Adaptive programming: object-based broadcasting . . . . .	43
2.4.4 Input mechanisms . . . . .	43
2.5 Discussion . . . . .	44
2.5.1 Programme delivery . . . . .	45
2.5.2 Devices . . . . .	45
2.5.3 Input mechanisms . . . . .	46

Table of contents	6
2.5.4 Navigation	46
2.5.5 Additional media activities	47
<b>3 Audition</b>	<b>49</b>
3.1 Introduction	49
3.2 Hearing: key principles	50
3.2.1 Physiology of the human auditory system	50
3.2.2 Spatial auditory perception	54
3.2.3 Masking	57
3.2.4 Auditory streams	59
3.2.5 Auditory attention	61
3.2.6 The influence of vision	62
3.3 Speech perception	64
3.3.1 Speech as a signal	64
3.3.2 Recognising words and understanding meaning	65
3.3.3 Concurrent speech	66
3.3.4 Influences of visual information on speech perception	76
3.4 Summary	76
<b>4 Auditory Display</b>	<b>79</b>
4.1 Serial speech displays	80
4.2 Non-speech auditory displays	82
4.2.1 Audification, parameter mapping and model-based sonification	83
4.2.2 Auditory icons	84
4.2.3 Earcons	85
4.2.4 Musicons	88
4.2.5 Spearcons	89
4.2.6 Auditory scrollbars and Spindex	91
4.2.7 Discussion	92
4.3 Concurrent auditory displays	93
4.3.1 Concurrent earcons	94
4.3.2 Concurrent auditory icons	97
4.3.3 Concurrent speech displays	99
4.4 Summary	110
<b>5 Menu navigation</b>	<b>113</b>
5.1 Introduction	113
5.2 Display design	115
5.2.1 Number of talkers	115
5.2.2 Spatial configuration	116

5.2.3	Interaction design . . . . .	116
5.2.4	Implementation . . . . .	119
5.3	Pilot Study . . . . .	120
5.3.1	Methodology . . . . .	120
5.3.2	Findings . . . . .	124
5.4	Experiment Proper . . . . .	125
5.4.1	Methodology . . . . .	125
5.4.2	Results . . . . .	131
5.4.3	Discussion . . . . .	138
5.5	Summary . . . . .	143
<b>6</b>	<b>Orchestrated synchronous companion experiences: Non-visual design considerations</b>	<b>145</b>
6.1	Introduction . . . . .	145
6.2	Attention with visual secondary programme content . . . . .	147
6.3	Auditory presentation of secondary programme content . . . . .	151
6.4	Design . . . . .	152
6.4.1	Content representation . . . . .	153
6.4.2	Concurrency of presentation . . . . .	155
6.4.3	Talker/performance factors . . . . .	160
6.4.4	Interaction design . . . . .	161
6.4.5	Notification design . . . . .	162
6.4.6	Spatial location . . . . .	164
6.4.7	Personal or shared experiences . . . . .	167
6.5	Secondary content applications . . . . .	169
6.6	Summary and outstanding questions . . . . .	170
<b>7</b>	<b>The effects of secondary content modality, location and the modification of programme soundtrack on television user experience</b>	<b>173</b>
7.1	Introduction . . . . .	173
7.2	Experimental design . . . . .	174
7.2.1	Independent variables . . . . .	174
7.2.2	Dependent variables . . . . .	175
7.3	Methodology . . . . .	179
7.3.1	Experiment structure . . . . .	179
7.3.2	Experiential limitations . . . . .	180
7.3.3	Prototype system . . . . .	181
7.3.4	Stimuli . . . . .	183
7.3.5	Pilot Study . . . . .	187
7.3.6	Methodological changes following the Pilot . . . . .	191

7.4	Experiment proper . . . . .	192
7.4.1	Demographics . . . . .	192
7.4.2	Procedure . . . . .	192
7.4.3	Experiment set-up . . . . .	193
7.4.4	Data treatment . . . . .	195
7.4.5	Quantitative Results . . . . .	197
7.4.6	Qualitative Results . . . . .	204
7.4.7	Discussion . . . . .	218
7.4.8	Limitations and future work . . . . .	228
7.5	Summary . . . . .	230
<b>8</b>	<b>Conclusion</b>	<b>233</b>
8.1	Introduction . . . . .	233
8.2	Menus . . . . .	235
8.3	Orchestrated synchronous companion experiences . . . . .	237
8.4	Concurrency within consumer interfaces . . . . .	239
8.5	Implications for the hypothesis . . . . .	241
8.6	Further work . . . . .	242
	<b>Appendices</b>	<b>245</b>
<b>A</b>	<b>Publications</b>	<b>245</b>
A.1	Onset Asynchrony in Spoken Menus . . . . .	245
A.2	Television and Additional Media Activity: A Taxonomy . . . . .	254
<b>B</b>	<b>Menu navigation experiment documents</b>	<b>273</b>
B.1	Ethics Documentation . . . . .	273
B.1.1	Pilot Ethics Documentation . . . . .	274
B.1.2	Main experiment Ethics Documentation . . . . .	283
B.2	Script . . . . .	293
B.3	Consent form . . . . .	297
B.4	Instructions . . . . .	298
B.5	Debriefing form . . . . .	300
B.6	Data access . . . . .	301
<b>C</b>	<b>Upmixing formula</b>	<b>303</b>
<b>D</b>	<b>Secondary Programme Content Experiment Documents</b>	<b>305</b>
D.1	Ethical approval . . . . .	305
D.2	Ethics application . . . . .	306
D.3	Consent For the main experiment . . . . .	316



D.4	Instructions before training . . . . .	318
D.5	Instructions before experiment trials . . . . .	319
D.5.1	Front group . . . . .	319
D.5.2	Side group . . . . .	319
D.5.3	SD-A group . . . . .	320
D.5.4	SD-V group . . . . .	320
D.6	Debriefing . . . . .	321
D.7	Response application screenshots . . . . .	321
D.8	Data access . . . . .	327
	<b>Glossary of Terms</b>	<b>329</b>
	<b>Abbreviations</b>	<b>331</b>
	<b>References</b>	<b>333</b>



# List of Figures

1.1	Comparison of (a) an early television set (adapted [cropped and resized] from van Beem (2013) — Licensed as CC0 1.0 Universal Public Domain Dedication) and (b) a modern day interface. . . . .	22
2.1	Categories of devices used for television viewing . . . . .	32
2.2	Illustration of manual and search-based navigation of hierarchical structures . . . . .	35
2.3	Examples of different television menu types . . . . .	36
2.4	Taxonomy of additional media activities relative to a television programme (Hoare & Hinde, 2016, p. 13) . . . . .	41
3.1	Annotated diagram of the human ear indicating the outer, middle, and inner ear sections. Adapted from (Plack, 2014, p. 54) . . . . .	51
3.2	Annotated cross-section of the cochlea. Adapted from (Plack, 2014, p. 56) . . . . .	52
3.3	Diagram showing the median, transverse and frontal planes. Adapted from (Blauert, 1997, p. 14). Locations are referred to in (azimuth, elevation) format. . . . .	54
3.4	Illustration of the additional path taken by sound to reach the contralateral ear for sources away from the median plane. Adapted from (Rumsey, 2001, p. 22) . . . . .	55
3.5	Spectrogram of speech signal containing the phrase “ran fast” showing the approximate location of the phonemes . . . . .	65
3.6	Visual representations of different stream segregation failures that may occur with speech. . . . .	66
4.1	Representation of McGookin & Brewster’s multi-modal display showing the focus and priority zones. The shade of the zones represents the increased importance required for a source to be sonified further from the focus. (Adapted from McGookin & Brewster (2001, p. 3)) . . . . .	96
4.2	Visual representation of the ‘Audio Hallway’ display. Adapted from Schmandt (1998, p. 167). . . . .	101
4.3	Visual representation of the ‘Dynamic Soundscape’ display. Adapted from Kobayashi & Schmandt (1997, p. 167). . . . .	102
4.4	Visual representation of the virtual dial display. Adapted from Frauenberger & Stockman (2006, p. 143). . . . .	103
4.5	The reduction in the number of concurrent sources when an onset asynchrony of 50% or greater is used (adapted from Ikei <i>et al.</i> (2006, p. 192)). . . . .	108

5.1	Illustration of the sliding window and segmented interaction methods . . . . .	117
5.2	Illustration of menu display design concept . . . . .	118
5.3	Control and function mappings used for the prototype display. . . . .	120
5.4	The visual display that presented participants with the target word and the current state of the interface . . . . .	123
5.5	Diagram of the experiment's running order . . . . .	130
5.6	The room setup used in the experiment . . . . .	132
5.7	Marginal means of the square root transformed total task durations . . . . .	133
5.8	Marginal means (original scale) for the number of trials requiring one or more repeats during each block . . . . .	134
5.9	Marginal means for the unweighted TLX scores . . . . .	135
5.10	Visual representation showing that the critical phonetic information for the first two words of a triplet with onset asynchrony . . . . .	139
6.1	Multiple resource model of attention (adapted from Wickens (2002, p. 163)) .	148
6.2	Proposed flow diagram of user interaction with ASPC . . . . .	161
6.3	Illustration of a shared experience comprising individual and shared streams .	169
7.1	Diagram of the information passed between the individual elements of the experiment system . . . . .	182
7.2	Example of the format used for the rating scales of the Likert-style questions in the response application . . . . .	183
7.3	Notation of the secondary content notification earcon . . . . .	184
7.4	Timing of the secondary content notification and secondary programme content	185
7.5	Diagram of the main programme content soundtrack manipulations used . . .	186
7.6	The four images shown to participants before trials to indicate how the SPC would be presented. . . . .	188
7.7	Diagrams of experimental set up . . . . .	194
7.8	Panorama of the set-up used for the experiment . . . . .	195
7.9	Plots of the 20% trimmed means of the ratings of disruption for the SPC on the experience of: the MPC; the MPC audio, and the MPC visuals. Plots of the 20% trimmed means of participant difference scores between the MPC treatments are also displayed for the ratings of disruption to the MPC and MPC audio . . . . .	199
7.10	Plots of the 20% trimmed means of the ratings of disruption caused to the experience of the SPC by the: MPC; MPC audio; MPC visuals. Plots of the 20% trimmed means of participant difference scores between the MPC treatments are presented for the ratings of disruption caused by the MPC and MPC audio . . . . .	201
7.11	Plots of the 20% trimmed means of the ratings of mental demand; temporal demand and physical demand. Plots of the 20% trimmed means of participant difference scores are also provided for ratings of mental and temporal demand	203

7.12 Plots of the 20% trimmed means of the ratings of effort, annoyance and preference. Plots of the 20% trimmed means of participant difference scores between the MPC treatments are also provided for the effort and preference ratings . . . . .	205
--	-----



# List of Tables

6.1	Tables of terms used to describe different components of a scheduled orchestrated companion experience . . . . .	146
6.2	Design considerations for scheduled orchestrated experiences in audio . . . . .	154
7.1	Scoring system, based on (Oldfield, 1971), used for calculating left and right scores for the laterality quotient. . . . .	179





# Acknowledgements

Firstly, I would like to thank those who have supervised me over the course of the project. I would like to express my thanks to my supervisors Mr. Tony Tew and Dr. Mike Evans for supervising me throughout this project. I am deeply grateful for the tireless support and guidance they have both provided me. I would also like to thank Prof. David Howard, who was also a supervisor for the majority of the project and provided valuable insights and interesting discussions at key points in the formative period of the research. Thank you also to Mike Armstrong, who stepped in to provide me with supervision for a period of my project, and with whom I had many interesting and influential conversations with.

I would also like to thank everyone working at BBC Research & Development North Lab, who hosted me for the majority of my research. Throughout my project many provided insights on my work, offered their assistance and expertise, volunteered to participate in my experiments, and made me feel like a welcome addition to the department—thank you to you all. I am particularly grateful the User Experience Research group for having me as a member of the team—it has been a privilege to work with such a talented group of people—and to the Audio team for their technical support. I am grateful to David Marston, for advising me on upmixing, providing me with his tool, and allowing me to report its workings.

Thank you to Charlotte Hoare for the many discussions throughout the project and for working with me on the development of the taxonomy of additional media activity and the development of the disruption rating scales.

I am thankful to the members of the Audio Lab in York, who always made me feel at home whenever I visited and accommodated my demands on their facilities. In particular, I would like to thank Andrew Chadwick, who provided valuable logistical and practical support.

I am grateful to Francis Duah, from the University of York Maths Skill Centre, for all of the advice that he gave on statistical methods.

Thank you to Prof. Rand Wilcox for the use of his R functions and updating the functions for me.

I would like to thank all of those who volunteered their time for this project either to be recorded, or to be participants in the experiments.

To my family, I am deeply grateful for all of the support you have given me over the years. Last but not least, thank you to Annie, for being there throughout all of it with encouragement, for listening to my ramblings and keeping me grounded. I could not have done this without her.

# Declaration

Parts of the research presented in this thesis have been published in:

Hinde, A. F., Evans, M., Tew, A. I. & Howard, D. M. (2015), ‘Onset asynchrony in spoken menus’, in: ‘Proceedings of the 21st International Conference on Auditory Display (ICAD 2015)’, pp. 86 – 93.

Hoare, C., & Hinde, A. F. (2016). ‘Television and Additional Media Activity: A Taxonomy’, Technical report, [online], Available: [https://figshare.com/articles/Television\\_and\\_additional\\_media\\_activity\\_A\\_taxonomy/3856164](https://figshare.com/articles/Television_and_additional_media_activity_A_taxonomy/3856164)

Copies of these papers can be found in Appendix A.

The development of the taxonomy for additional media activities and the disruption rating questions (discussed in Section 7.2.2) were developed in collaboration with Charlotte Hoare (University of Bath). The taxonomy was developed with equal contributions from both parties. With the disruption scales, the identification of the factors and general design of the ratings came through equal contributions from both researchers. The wording of the questions that were used within this thesis was principally developed by Alistair.

All other aspects of the thesis are my own work and all other contributions are explicitly stated or referenced. This thesis has not been submitted for any other award at this, or any other, institution.

This research was funded by an EPSRC ICASE award with BBC Research & Development and The University of York for the project “Non-visual Displays for Connected Television”.



# Chapter 1

## Introduction

### 1.1 Overview: The dominance of vision in television displays

Throughout human history, people have processed important information from their environment through the senses of sight, touch, smell, taste and hearing. The ability to draw information from these senses played a vital role in enabling our ancestors to survive their natural environments. In the present day, we spend increasing amounts of time accessing information from man-made digital environments or experiences. Much of this information, however, is accessible only via visual displays and, as technology becomes increasingly ubiquitous, we are confronted with ever more screens and graphical user interfaces (GUIs) imposing higher and higher demands on our visual attention. Meanwhile, the capabilities of our other sensory modalities are exploited much less.

The development of GUIs has been a key contributing factor in enhancing the usability of computers and digital systems for the general public. From a manufacturer's perspective, the GUI also provides customisable control systems without the expense of custom hardware design, development and implementation. GUIs also allow a large number of functionalities to be achievable within one machine, the personal computer (PC) being one of the most apparent examples of this. A GUI's usability, however, is often dependent on a user's ability to visually attend to the interface. This, combined with the shift away from single-function interfaces, has made many of these systems more difficult to use non-visually. One technology for which this trend has been clearly observed is television.

Over the past few decades, the face of television has changed dramatically from simply allowing passive viewing of a small number of broadcast streams to access to hundreds



**Figure 1.1:** Comparison of (a) an early television set (adapted [cropped and resized] from van Beem (2013) — Licensed as CC0 1.0 Universal Public Domain Dedication) and (b) a modern day interface.

of channels with interactive content. In recent years, manufacturers have added internet connectivity, either by including network connectivity on television sets, or via network-connected peripherals (e.g., games consoles and Blu-Ray players). This connection to the internet allows users to install applications, view internet-based on-demand content and browse the web.

This explosion of content has required the development of increasingly complex user interfaces. A comparison of early televisions controls, which comprise a few buttons or dials for tuning the station and controlling the volume, to those found on one of today's televisions, highlights the escalation in complexity (see Figure 1.1). This illustration is, however, only part of the story. Additionally, television interfaces are no longer restricted to programme selection and device configuration. With the addition of network connectivity, televisions may communicate with other connected devices over the home network and present synchronised content from other devices' displays to facilitate new user experiences that were not previously possible. The exploration of the capabilities of these developments has largely remained centred on the introduction of additional screens. For many users, most of the time solely presenting television interfaces as GUIs provides adequate access to the desired services. In scenarios where it is not possible for the user to attend visually to the screen, it becomes very difficult to navigate these systems or access many of their functions. It is important to state that this is not a call for a reversion to the restricted functionality of past systems, but rather a demonstration that other, non-visual approaches need to be developed to allow access to these modern services and to ensure that new non-visual experiences are also considered. Due to the dominance of visual displays in computer interfaces, there has been extensive research

into various elements of their design. Making use of the user's other senses, such as hearing, touch and smell/taste in human-computer interaction (HCI), however, has not been explored to the same extent. This, therefore, makes the task of non-visual interface design difficult for developers in the commercial sector, who are faced with limited scientific evidence to support alternative approaches.

## 1.2 Users effected by visio-centric design

So far it has been mentioned that a user may have difficulty using these systems if they are unable to access the display visually. Some consideration will now be given to user-groups who are likely to be affected by current visio-centric design practices and who may benefit from the development of non-visual displays.

Perhaps the most obvious demographic likely to experience difficulties with visual-only displays are the users with visual impairments or who are blind. In 2008, it was estimated that there were 1.8 million partially sighted or blind people in the United Kingdom (UK), with this figure predicted to more than double to four million by 2050 (Access Economics Pty Limited, 2009). Though it is tempting to think of television as a predominantly visual medium, research surveys have indicated that visually impaired users also consume considerable amounts of television (Pettitt *et al.*, 1996; Woods & Satgunam, 2011). It is clear, therefore, that the greater development of non-visual alternatives would potentially benefit this sizeable group.

Difficulty with visual displays is not solely a concern for those who are unable to see the content on the screen. Users may struggle with textually presented information due to their educational level, learning disabilities and cognitive disabilities (Gribbons, 2008). One method of making these systems more accessible for users with low literacy is the inclusion of additional icons and/or images (e.g., Götze & Strothotte, 2001; Shakeel & Best, 2002; Medhi *et al.*, 2007a, b). For some forms of information, however, suitably clear graphical representations may be impossible to find. This has led some researchers to suggest the use of alternative modalities, most commonly audio (e.g., for severe dyslexia (Dix *et al.*, 2004) and for illiteracy (e.g., Huenerfauth, 2002; Medhi *et al.*, 2007a; Knoche & Huang, 2012)).

It is important to recognise that even those with normal vision are sometimes unable to visually attend to a screen. Sears *et al.* (2003) described such scenarios as situationally-induced impairments and disabilities (SIID). In the context of television, SIID

could occur when the user is engaged in another task requiring visual attention, or due to the limitations of the device that they are using (e.g., insufficient screen space on a smartphone). The addition of non-visual displays to television experiences can therefore be seen as following the principles of *universal design*, which is defined by Dix *et al.* (2004, p. 366) as “the process of designing products so that they can be used by as many people as possible in as many situations as possible”. This project will, therefore, consider the design of displays for a general population, rather than solely for users who have visual impairments or who are blind. Though this may appear to be a purely philosophical consideration, it may have important implications for the approaches to be considered in this thesis. Many solutions that have been developed for accessibility purposes are based on the premise of a highly-trained user (e.g., braille and sign language). When considering the general population, however, it should not be assumed that users are sufficiently familiar with these techniques. Given the use case of television, it is also unlikely that users will be willing to go through the same amount of training as many accessibility systems require. The display must therefore use common, or easily learnt, means of communicating its information.

There is some evidence that standards organisations are beginning to realise the importance of non-visual equivalence for television interfaces. The lack of non-visual accessibility for on-screen TV menus and options has been highlighted as an area in which development is needed for the improvement of TV access in the UK (European Blind Union (EBU), 2008). Its demand is such that the U.S. Government have introduced rules that stipulate the use of audible equivalents in on-screen menus (U.S. Government, 2013a) and navigation devices for programme guides (U.S. Government, 2013b) for users who are visually-impaired or blind. With this in mind, it is expected that this will become an increasingly important area of interest for TV/set-top box (STB) manufacturers and broadcasters.

Non-visual displays, however, should not be thought of solely as alternatives when the visual consumption of information is inappropriate. It often appears that the default behaviour of user experience designers is to produce visual displays with little or no consideration of the potential benefits of using other modalities. There is value in considering the potential of non-visual approaches for future experiences where visual presentation may not be the optimal solution. This could be due to the types of information being represented or to other demands being placed on the users visual attention. For this reason, this research is not restricted to issues of equivalence but also considers the capabilities of non-visual perception and how these can be utilised to enhance user experiences.



## 1.3 Auditory display

Audition, or the sense of hearing, is a powerful sense both for communication and for providing information about our surroundings. Spoken language has been the primary means of communication between people throughout our history. Furthermore, the development of music demonstrates the communication of information from composer and performer to the listener through non-vocal sound. Audition, therefore, appears to have great potential in the development of non-visual displays to communicate information from a computer interface to a user.

The inherent temporal aspect of sound is, however, problematic. Whereas in visual displays a user may skim over much of a visually-complex scene or divide their attention between several concurrent sources, auditory displays are generally sequential, which has ramifications in terms of the speed and timeliness of information access. Conversely, real-world auditory events are seldom isolated, but occur alongside a mixture of other sounds. Cherry's (1953) 'cocktail party problem', whereby one understands a talker despite the presence of other conversations, is one notable example of the human ability to segregate source sources. It suggests that some concurrency may be possible within the design of auditory displays. This research seeks to investigate the use of concurrency within use cases for connected television.

## 1.4 Hypothesis

Throughout the course of the work it is intended to explore the design of auditory displays for connected television use cases. The project will focus predominantly on the potential of concurrency within consumer systems and how this affects measures of performance such as speed and accuracy, as well as experiential aspects of the displays.

The work will explore the hypothesis that concurrent auditory displays can:

1. facilitate faster navigation of menus without negatively affecting accuracy or user experience;
2. provide less disruptive and demanding display of additional secondary programme content than serial alternatives.

## 1.5 Contributions

This research provides several contributions to the field of study in addition to the stated hypothesis.

This work comes at a time when television experiences are rapidly changing. This thesis provides a review of some of these trends and suggests taxonomies and terminologies by which these new experiences may be referred to. While serving primarily to aid in discussions within the thesis, it is hoped that this will help the practitioners looking to create new media experiences as well as the researchers attempting to discuss them.

The work also outlines the design of prototype displays for two television use cases. The discussion of these design decisions is intended to inform designers of auditory displays for use cases that present similar sets of challenges.

In the first of the use cases, considering the representation of menus through auditory display, the experimental work attempts to separate the effects of temporal overlap and onset asynchrony on navigation speed and accuracy in auditory menus. This is an important consideration in displays of this type and it is hoped that this work will inform designers considering concurrent spoken menus and researchers looking at the perception of concurrent speech.

The second of these two use cases, the presentation of auditory secondary programme content alongside a television programme, is a novel proposition. It should be clear that while attempts are made to draw insights from other research in related areas, this work represents the first exploration of this area. This initial discussion of auditory secondary programme content is, therefore, intended to serve as a catalyst for those considering the design and development of new television experiences, from both academic and industrial backgrounds. Furthermore, in considering these experiences it is clear that studying this use case provides a new platform from which to explore the limitations of human attention and perception.

In order to study the display variations considered in this thesis, new methodologies are developed and presented. It is hoped that the methodologies adopted in this work will assist future researchers exploring scenarios with similar design challenges.

## 1.6 Structure of thesis

The first three chapters of this thesis are dedicated to providing a theoretical background to the project and a review of the existing relevant research material. Chapter two provides a more in-depth introduction to the state-of-the-art in television interfaces. The chapter introduces taxonomies used throughout the project and explains restrictions on the scope of this project. Chapter three provides an introduction to the human auditory system and discusses both physiological and psychoacoustic principles. Special attention is paid to the perception of speech, particularly in instances in which multiple talkers are present at the same time and the factors that are important in these scenarios.

Chapter four provides an introduction to auditory displays. It begins by outlining alternative non-visual approaches and the reasons for focusing on auditory displays. The chapter goes on to review previous work on auditory displays, focussing initially on serial approaches before moving on to concurrent auditory displays. Previous approaches are discussed in relation to the potential application of these methods within connected-television displays.

Chapter five focuses on the case of menu-navigation and providing faster non-visual navigation through the use of concurrency. An interactive system is proposed and the considerations associated with the design process are outlined. A pilot study is presented followed by a larger experimental evaluation of this system. The results from this study are discussed and implications identified for the use of concurrent speech.

Chapter six focuses on the second use case, orchestrated synchronous companion experiences. A discussion of previous work on visual companion experiences is provided and work on audio description is considered and used to inform the design of an auditory display for secondary programme content. Chapter seven then presents experimental work (a small pilot followed by a larger study) on the effects of different auditory and visual presentations of additional content when the main programme's soundtrack is manipulated or left unaltered on disruption, workload and preference. The findings are used to consider the potential for auditory companion experiences for television and the part which concurrency may play within them.

Chapter eight considers the findings from both use cases and discusses the implications they have on the hypothesis. These findings are used to provide guidance for other researchers within this field and to suggest further work which falls beyond the scope of this project.



## Chapter 2

# Television experiences: Now and the near future

### 2.1 Introduction

Over its relatively short history, television has seen considerable changes both in terms of the technologies that are used to distribute and consume content, and the formats of the content itself. From black and white analogue video on a few channels distributed by terrestrial broadcast, to high-definition, colour, digital video with a plethora of channels and a variety of broadcast methods (e.g., terrestrial, cable and satellite). These technological developments have also impacted the user experiences of television. In some cases, the effects are clear (e.g., the provision of colour images), but in others it is less obvious, such as broadcast methods. The change from analogue to digital broadcasts allowed broadcasters to deliver more channels, provide additional programme guides and interactivity (Digital UK, 2012). The user experience and the technology are, therefore, fundamentally intertwined.

Over recent years, the roll-out of high speed internet networks and the convergence of media playback and personal computing technologies is once again changing television. Over-the-top (OTT) services (e.g., BBC iPlayer, Netflix and Amazon Prime) are providing television programmes as on-demand pieces of content, rather than streams of broadcast material. Reports from the UK are showing increasing use of these services (Ofcom, 2016). The popularity of these services has led some to declare that the death of broadcast television is a prospect for the near future (Hastings quoted in Hecht (2014)). These new services still provide television content and, by extension, television experiences. This is certainly not the

death of television. Instead, these technologies are facilitating content providers to expand the definition of television user-experience to encompass a larger range of devices, services and use-contexts.

This chapter provides a snapshot of the current state of television and emerging trends in programme distribution, the devices used to consume televisual content, common interfaces and experience elements for television. From this, key features are identified that are explored within the context of auditory display throughout the rest of the thesis.

## 2.2 Programme delivery

While the mechanisms of programme delivery are largely hidden from the user, the method used has important implications for the experiences that may be created. This section, therefore, intends to give an overview of the methods and their associated limitations.

Traditionally, broadcast television distributed streams of channel information as radio signals from ground-based transmitters (terrestrial), via cable, or satellites (Ibrahim & Trundle, 2007). More recently, with the spread of high speed broadband connections, distribution methods based on internet technologies have emerged. These internet-based methods can be split into two categories: internet-protocol TV (IPTV) and OTT or internet TV. IPTV refers to systems in which content is delivered to subscribers via closed networks that are run by the content provider (Montpetit *et al.*, 2011). Current examples of this type of service in the UK are offered by Virgin Media, BT, TalkTalk and Sky (USwitch, n.d.). Montpetit *et al.* (2011, p. 521) describes the broad, alternative category “internet video”, which encompasses all video delivered over the public internet. This project, however, will focus on only professionally produced “TV” content. OTT services offering TV content include video on-demand (VOD) services such as Netflix, Amazon Prime, Channel 4’s All4 and BBC iPlayer.

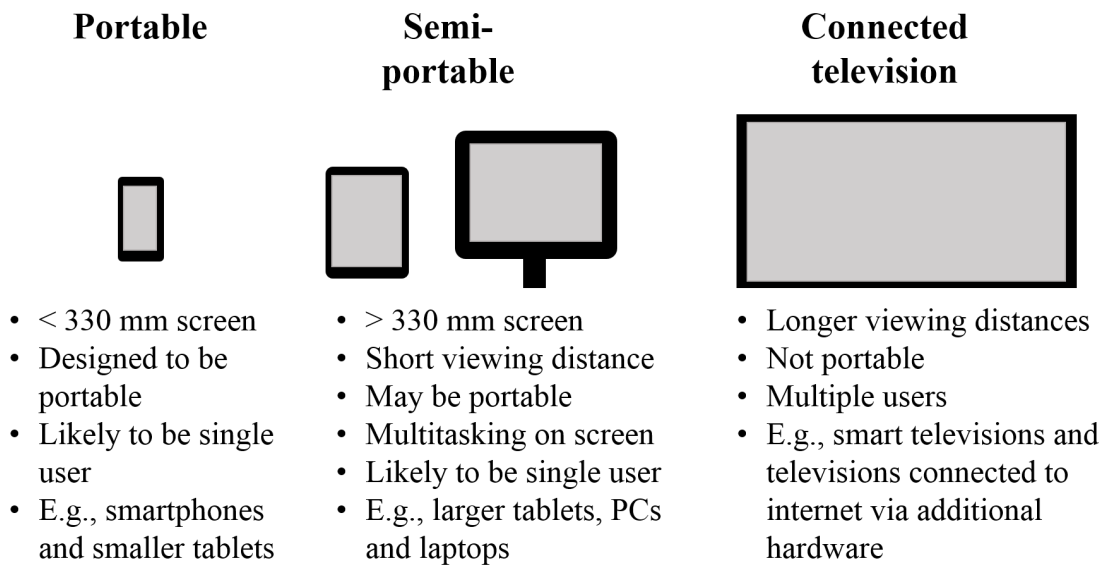
Internet delivery offers several advantages over traditional broadcast methods. Firstly, traditional broadcast mediums have been restricted in the amount of content they can transmit due to the bandwidth restrictions of the channel. In the case of terrestrial broadcast, radio frequency (RF) spectrum is a finite, valuable and tightly-regulated commodity. As a result of this, there is little scope for including any supplementary content alongside a main programme. Additional server space, however, is comparatively cheap and simple to acquire. This means that extra features, such as alternative coverage of an event, can be delivered with much lower overheads. Secondly, internet-based delivery of some, or all, of the content offers a

much more versatile experience than traditional broadcast methods. As an increasing number of devices can now connect to the internet it is possible to view content on a wider range of devices, and consequently in a larger variety of contexts than with traditional broadcast methods (discussed further in Section 2.3).

Broadcasting has historically been a largely one-way system. The content provider encodes and transmits information which the user's device receives and displays. Incremental steps towards interactive television have been made by many technologies throughout the history of television broadcast (Jensen, 2008). While digital television provides some interactive experiences, these usually consist of choosing which page of text or programme stream being received by the user's device is presented (Vinayagamoorthy *et al.*, 2012). Through making use of an internet connection, however, it is possible to establish a two-way connection between the user and content provider, opening up the potential for interactive television experiences. This two-way connection can allow users to request specific content through the use of VOD services and perform web-style interactions (Vinayagamoorthy *et al.*, 2012).

## 2.3 Television devices

Traditionally, television sets have been set apart from other display devices by their ability to receive terrestrial, satellite or cable broadcast streams. Television experiences, however, are no longer constrained to devices capable of receiving traditional broadcast signals. Due to the emergence of internet-based delivery methods, discussed in the previous section, any device with internet connectivity is capable of accessing television content. Furthermore, technological convergence has meant that devices originally associated with other tasks (e.g., PCs, phones, watches) are now also capable of playing digital video. Television experiences can, therefore, now occur on a plethora of new devices. This has both technical and experiential implications. Unlike the traditional television set, these devices are not designed with the sole purpose of viewing television programmes. Some issues which may emerge due to this include the amount of available screen space, either due to screen size or the window used for video playback, and comparatively poor audio quality due to speaker arrangement and size. The portability and multi-functionality also means that the contexts in which the content is consumed are much more varied than the traditional model of television viewing, which was restricted to the living-room (e.g., using a smartphone or tablet on public transport, or on a laptop whilst browsing the web in another window). Given that such a wide range of devices and, by extension, experiences are possible within this future vision of



*Figure 2.1: Categories of devices used for television viewing*

television, some classifications are introduced to aid the discussion (see Figure 2.1).

The first class of device we consider is the ‘portable’ device. These are multifunctional devices with an internet connection. They are designed to be portable and therefore have restricted screen sizes (up to  $\approx 330$  mm) (e.g., smartphones and most tablets). The portability of these devices means that they may be used outside the living room in a wide range of environments, ranging from other rooms within the home to outside or on public transport. While it is possible to have shared viewing experiences with small groups, it is assumed that these devices are primarily used by one user at a time. With operating systems introducing the ability to open more than one application window at once (Android, n.d.; Apple Inc., n.d.a), it is possible that a user may be also interacting with other content in a separate window on the same screen. Where this is the case or the device has a smaller screen, however, the limited screen space is an important factor. There is also some indication that these devices may also be mostly associated with specific formats, with Ofcom reporting that “users are twice as likely to use their phones to watch short-form video clips than for streaming television programmes or films” (Ofcom, 2015, pp. 6-7).

The second class of device is the ‘semi-portable’ device. These multifunctional devices offer larger screen sizes than portable devices (upwards of 330 mm), but are designed for comparatively short viewing distances (e.g. larger tablets, laptops and PCs). Similarly to portable devices, it is likely that most of the usage is single user. The context of use is more likely to be within domestic environments, although they may also be used in the varied



environments mentioned for the portable devices. Screen space is less limited than with the portable devices, although multi-tasking on the same screen is also more likely, which may restrict the screen space available for the display of video content and make it similar to some of the portable devices.

The third class of device is the connected television. This refers to a system designed for viewing distances in excess of one metre and with internet connectivity. Ofcom (2013) refer to two types of television that connect to the internet: *smart televisions* which have integrated internet connectivity, and *internet-enabled televisions* which have internet connectivity due to an additional piece of hardware (e.g., games console or STB). Within this thesis, connected television is used in the same manner as the BBC (2013b), as an umbrella term to refer to both types of device. Due to the size of these devices, it is assumed that they are restricted to indoor use, most typically within the living room. Unlike the other classes of device, the connected television is likely to be part of shared experiences. These devices may also receive content from traditional broadcast methods, though this is not a requirement within this work.

Perhaps the most interesting aspect of the connected television device, is the fact that it is connected to a local home network. While all of the previously described device classes have access to content via internet protocol (IP), the connection to a local network means that the television can communicate with other devices on the network. This communication may allow these connected devices to act as input devices or additional display outputs for the television. This project views connected devices as elements within the television system and considers their display capabilities as an extension to the main television set. This opens up a host of creative possibilities for display and interaction design that have yet to be fully explored.

## 2.4 Elements of the television experience

When considering TV experiences, it is easy to focus solely on the programme content. A user's television experience, however, may also consist of searching for new content, accessing additional information through interactive services and modifying device/service settings. This section discusses four key aspects of television interfaces and additional services that are currently provided, or are likely to become common in the future: interfaces for navigation and choice, additional media activities, adaptive programming, and input mechanisms.

### 2.4.1 Navigation and choice

Interfaces that allow users to navigate options have been integral components of the television experience for as long as there has been more than one broadcast television station. With digital broadcast, internet connectivity and ever more capable TV sets and STBs, the amount of choice available to users has increased dramatically (e.g., channels, apps and VOD programmes). Without proper design the process of searching for and choosing the best option may be off-putting for users, leading them to settle for less desirable choices or avoid using these features at all. The navigation of options has long been a problem for interface designers working on PCs. As these technologies have converged, common features of PC interfaces have been applied to the television problem space. Within current television interfaces, menu and search-based interfaces are now commonplace.

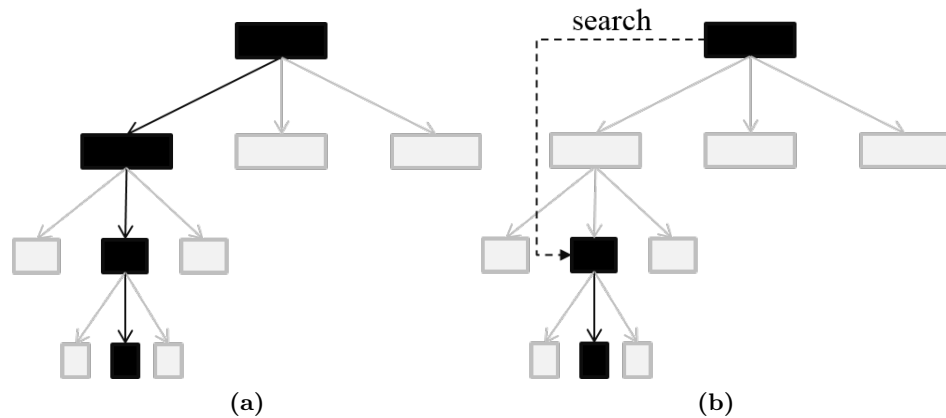
Menus are a long-standing feature of GUIs to such a degree that Norman (2008, p. 556) remarked:

“if Shakespeare could write “all the world is a stage,” an interface designer could point to the computer screen and say “all the interface is a menu.””

With the convergence of television and computing technologies, this is now also true of television interfaces. Within television systems, menus facilitate programme selection (for viewing or recording), selection of service applications on connected devices, navigation of interactive content options, and the configuring of device options. As the number of options increases, so does the prevalence of the menu.

The reasons for the initial popularity of menu systems becomes clear when comparing them to the alternative, command-based interfaces. Norman (1991) pointed out that menus present options to users, reducing the amount of training required and allowing them to be used by both novice and expert users. He contrasted this to command-based interfaces, which provide little information to assist users and therefore require a greater amount of training before use. Menus can take many different forms depending on the options that they represent. Norman (1991) outlines multiple different types of menus, the simplest of which is a stand-alone one-dimensional list of options, or *single* menus, while more complex multi-level variations impose some structure on the options. The *hierarchical* menus are the most extreme form of this structuring where users navigate through a tree-like structure, refining choices as they move down a specific branch (Norman, 1991).

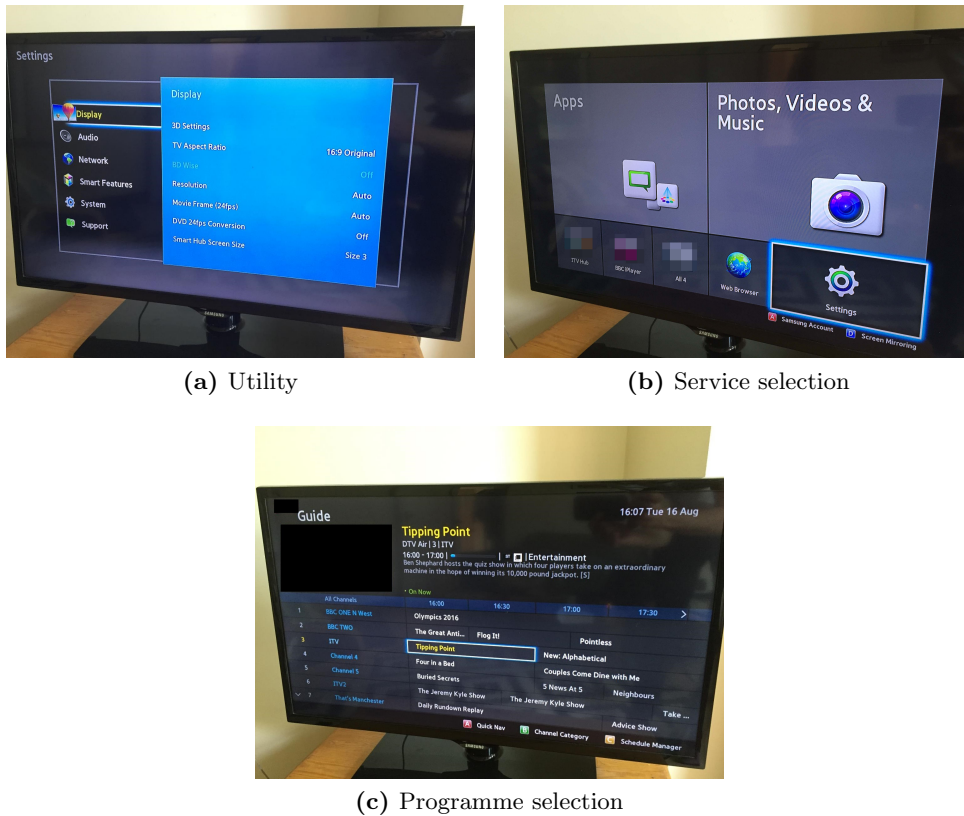
Menus can facilitate a range of different behaviours in users looking for content. Canter



**Figure 2.2:** Representation of hierarchical menu navigation (black arrows and items indicate the route): a) manual navigation of menu structure, b) navigation when a search function has been used to narrow results

*et al.* (1985) described five categories: *scanning*—navigating across many items paying little attention to the details of specific items, *browsing*—moving through options without any specific target until something is considered interesting, *searching*—navigation motivated to find a specific item, *exploring*—navigation through different parts of the menu structure to see the range of options available, and *wandering*—navigation in an unstructured fashion revisiting the same options. Search functions and natural language interfaces offer alternatives to menu interfaces. Depending on the sophistication of the implementation, these systems can help to reduce the need for users to remember specific commands or work out the exact location of desired options. These interfaces, however, only fully offer an alternative for users undertaking the behaviour referred to by Canter *et al.* (1985) as *searching* tasks. In these scenarios, a user has a strongly defined idea of a target. This means that, using well specified search terms, users can navigate directly to the desired option without any additional choices. In all of the other navigation behaviours described by Canter *et al.* (1985), these services will not completely circumvent the need for some menu navigation. Where the target of an item is either not, or is only partially defined, search functions may reduce the number of options that will be returned. Search functions can, therefore, be seen as a shortcut for traversing part of the menu structure. They do not remove the problem of menu representations within displays — at best they serve to allow accelerated navigation to the lower levels in the structures (see Figure 2.2) and at worst they limit exploration and discovery of new content.

Connected television systems may feature many types of menus, each with their own specific design considerations. With the ability to download applications on televisions offering a diverse range of experiences, there may be any number of different menus featured within



**Figure 2.3:** Examples of the different menu types on a Samsung UE32F5000 and Samsung BD-F6500 Blu-Ray player

third-party applications. There are, however, a few basic types which seem to be fundamental to connected television systems. *Utility menus* are those which present general configuration options such as picture settings and network connection settings (see Figure 2.3a). These are likely to be used fairly infrequently in the majority of cases, when first setting up the television in a new location. This means that on each use it is not possible to assume that users will have a strong memory of the menu contents or its structure. These menus are likely to be comparatively small. The items and their structure will generally remain constant, though occasional software updates may lead to slight variations.

*Service selection menus* offer users choices between different applications available on the connected television (see Figure 2.3b). These may be core applications provided as part of the television/STB operating system (OS) or third-party applications (e.g., VOD services and games). These menus are likely to be commonly used when users are switching between applications. They comprise a few options that remain fairly constant. In some cases, users may download additional applications. Even when this is the case, these additions will be infrequent.

The third type of menu is the *programme selection menu*, which allows the selection of

content to view or record (see Figure 2.3c). This menu differs considerably from the other two types. This category includes both the electronic programme guides (EPGs) used to navigate between broadcast channels and VOD selection interfaces. The content in these menus is likely to change regularly and offers a vast choice. Consider, for example an EPG, in some cases there are hundreds of channels to choose between, each offering different content throughout the day. Of the three categories of television menus discussed here, this type will see by far the heaviest use. In traditional digital television, unless a user wants to flick between individual channels or knows the number of a desired channel, they must traverse the EPG each time they want to select a new programme to watch. With VOD services, some form of programme selection menu must be navigated every time the user wishes to watch a different title.

Programme selection may display information regarding the duration, title, genre, time of broadcast and a synopsis. While it is likely that the majority of users will have a reasonably well defined target during utility and service selection menu navigation, this is not the case with programme selection menus. With EPGs it is common for users not to know exactly what programme or channel they wish to watch and instead have personal selection criteria, depending on the context (e.g., who they are viewing with, time restrictions, mood) (Elsweiler *et al.*, 2010). With this in mind, it is important that any display presents information regarding these factors so that users can decide what content best satisfies their needs.

With the uptake of IP content delivery OTT services are using web technologies (e.g., HTML5, and JavaScript) to create their programme selection interfaces. These can dynamically update, so that users are presented with different options on each visit, and respond to user interactions. For example, while entering a search term the displayed options may update for each added letter, converging on the desired subset.

A difficult-to-use utility menu is likely to have minimal effect on most of a user's television experiences, so long as they are still able to eventually locate and access the desired function. Due to the more regular use of service and programme selection menus, the ease with which they may be traversed will have a much larger impact on a user's experience of watching television. With the increasing amount of content available on televisions from an assortment of broadcast channels and online video services offered by a plethora of different companies, the user has a seemingly endless set of options to choose from. If the display is too slow or difficult to use, users may be put off fully exploring available options and miss out on content which they would have otherwise been interested in.

### 2.4.2 Additional media activities

Television experiences expand beyond navigating to and viewing television programmes. Benford *et al.* (2009) introduced the concept of trajectories of user experiences that pass through space, time, roles and interfaces. Additional media activities may be seen as extending or altering a user's trajectory for a specific television programme. These additional activities may occur on the device used for watching the television content or a host of other devices. Television sets may also be used to present non-programme content.

The concept of providing non-programme information as part of the television experience has been part of broadcast television for a long time. The first of these systems was the BBC's *Ceefax* teletext service which was broadcast from 1974 (BBC, 2013a). Ceefax provided users with information such as news, weather, sports scores, scheduling information, and lifestyle features (e.g., recipes and share prices) (Hand, 2012). Within the current digital broadcasting systems, teletext services are still available offering similar content with the addition of some new interactive features. Teletext services display predominantly textual information on the screen. In some cases, this completely obfuscates the picture, as with the early Ceefax system, or appears as a graphical overlay on top of the television material as with the current BBC Red Button interface.

With the spread of internet access, much of this information became available online via PCs. This development also saw the creation of websites which contained information about specific shows and forums in which fans could discuss the content. While originally most sites were created by the fan-community, the programme creators began to provide official sites (Gillan, 2011). Many shows now have official web-pages that provide additional content such as character profiles, mini-games, behind-the-scenes footage, interviews, further information on topics, or even in some cases with drama exclusive mini-episodes covering sub-plots.

As internet and traditional broadcast methods converge and television experiences spread across a wide variety of devices, the boundary between teletext and web-browsing experiences becomes harder to define. Most connected devices offer web browsers or third-party applications that allow users to access the same information that could be accessed through these systems. It seems, therefore, that systems like this that draw information from the internet could be viewed as no more than specialised browsers. It follows that these elements may actually be experienced on other devices that are not involved in presenting the television programmes. When viewing using a connected television, secondary portable devices may be used to access information on the internet, or may be used to display information passed

on from the connected television over the home network. Evidence indicates that this has become common, with multiple reports referring to people using other devices to access information while watching television (Consumer Electronics Association, 2014; The Nielson Company, 2013). This has been exploited by some content producers, who have released *apps* to allow the audience to receive additional content on mobile devices used alongside a programme experienced on the television. Experiences have been proposed to provide the user with further information on the programme subject through ‘curated’ collections of media (e.g., Basapur *et al.*, 2011; Jaye, 2012), to encourage game-like interaction (e.g., Luyten *et al.*, 2006; Williams, 2013), or social interactions (Basapur *et al.*, 2012) throughout the course of the programme. These systems can be synchronised with the programme either through the use of watermarks embedded in the content, direct messages sent to the second device from the broadcaster, or through sending time-coded messages from the primary device to the second device using a local network (2nd Screen Society, n.d.).

Much of the work in this field has considered applications which display information on small personal device screens. Recent work from BBC Research & Development (n.d.b), however, proposes a future in which users’ walls are used as elements of the television display using ‘smart wallpaper’. Such displays are intended to present additional information, increase immersion and allow the television experience to expand beyond the traditional confines of the screen (BBC Research & Development, n.d.b). Conversely, it has recently become possible to include curated interactive content within an on-screen programme through embedded video technologies (e.g., Touchcast, n.d.). These technologies allow the content producers to position additional content, such as polls, live feeds and related media, on the screen alongside the main content (Touchcast, n.d.). Though these tools are initially intended for PC or tablet use, with the increasing capabilities of televisions it is fully believable that such experiences will be possible on connected televisions in the near future. In some ways, this technology seems a natural progression from the ‘Red Button’, providing curated or user-selected content alongside the programme. Though this technology is new and interaction patterns for televisual experiences are yet to be defined, it would seem likely that this way of presenting content will become part of the television user experience of the future.

A technology has been developed by BBC Research and Development called ‘universal control’ (Barrett *et al.*, 2011a, b), which allows multiple devices to communicate with a STB over the home network. With little requirement for additional technology, a wide range of devices is available to control or to be controlled by the television content, extending the display

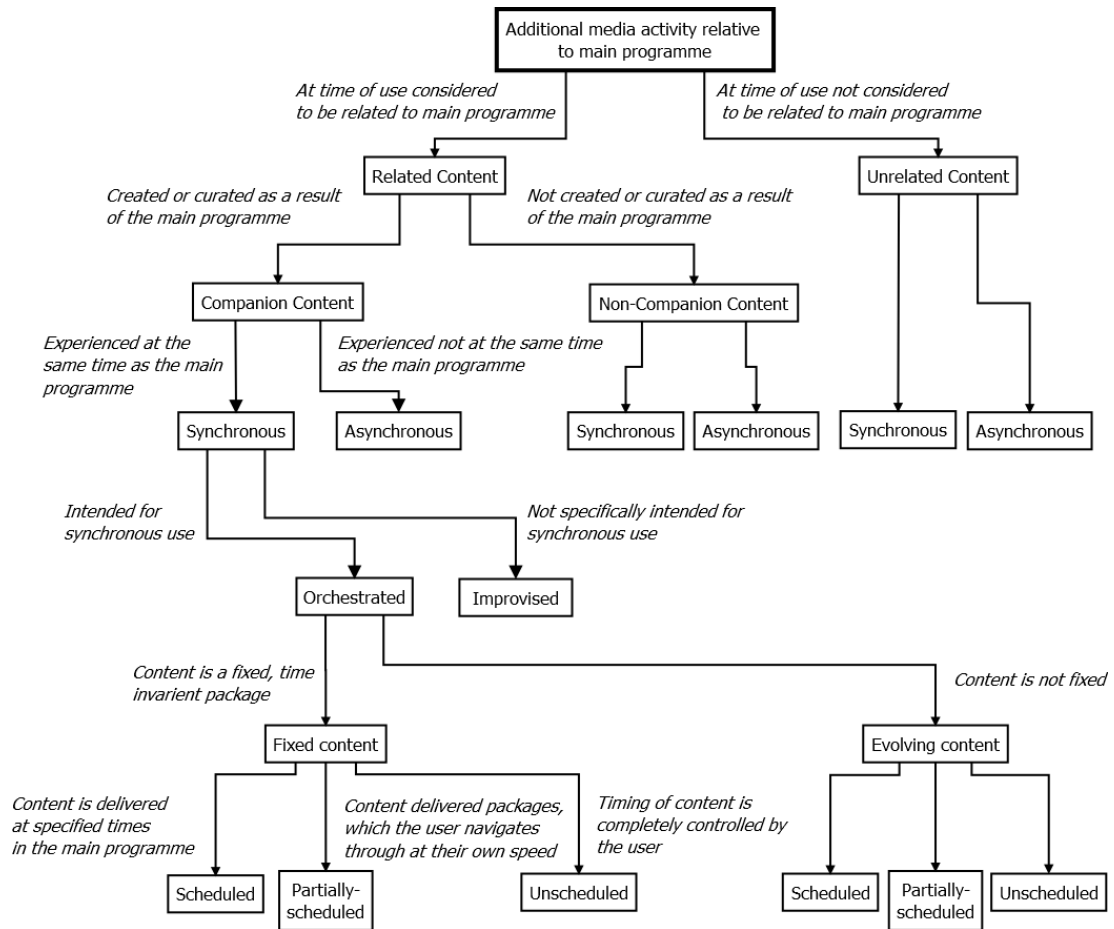
beyond a single screen and the interface away from the traditional remote control. Whilst the examples mentioned so far have all been screen based, universal control has the potential to be exploited for less conventional display technologies. Some researchers have been exploring the use of off-screen *tangible* objects to enhance the user experience. A notable demonstration of this was the modification of a toy character from the BBC show *Dr. Who* so that the toy would move and make sound effects at specific points in the show (Jolly & Evans, 2013).

The additional media activities that have been outlined in this section vary considerably in terms of the forms of interaction involved, the information they communicate and the context of their use. This project will follow the taxonomy proposed by Hoare & Hinde (2016) (included as Appendix A.2). This taxonomy is used to categorise all additional media activity in relation to a specific television programme. In line with the motivations of this thesis, the taxonomy does not include considerations of device or modality. It should also be noted that accessibility services are classified as part of the main programme and are not considered in this taxonomy. Classifications are based on the content that is presented, when it is experienced and the intention of the person who facilitated the experience. The taxonomy is presented as a tree-like structure to show the stages of classification (Figure 2.4).

The first distinction is made between content that the user considers to be related to the main programme at the time of experiencing it (*related*) and content which is not (*unrelated*). *Related* content is then further split based on whether it was created or curated as a result of the main programme, by either the makers of the programme or a third party. If this is the case, it is classified as *companion content* or, otherwise, *non-companion content*. To illustrate this distinction, consider episode ‘Cities’ from the BBC’s natural history programme Planet Earth II (Devas, 2016). To accompany the series and specific episodes there is a website (BBC, 2016) comprising a host of additional media to accompany the show, such as behind the scenes stories, footage, 360° videos and image galleries. This is an example of companion content because the content has been created, or at least curated, as a result of the show. This is contrasted with a general resource on a species featured in the episode, such as the Wikipedia page on Gray Langurs (Wikipedia Contributors, 2017). This would be considered as non-companion content because, despite the fact the content is related to the main programme, it has not been created or curated as a result of the main programme.

The taxonomy then distinguishes between content that is experienced at the same time as the programme (*synchronous experience*) and content experienced at any other time (*asynchronous experience*). This categorisation applies to companion, non-companion and





**Figure 2.4:** Diagrammatic representation of the taxonomy for additional media activities relative to a television programme (Hoare & Hinde, 2016, p.13)

unrelated content.

Synchronous experiences are then split based on whether the experience was intended for synchronous use with the programme (referred to as an *orchestrated experience*), or if it comprises a general purpose resource selected by the user (referred to as an *improvised experience*). Orchestrated experiences are then categorised depending whether content is a time-invariant package (*fixed*) or whether it includes elements that may change (*evolving*) such as most popular tweets on a hash tag at the time of watching. The amount that the orchestrator has specified the timing of content is the basis of the next categorisation. Synchronous companion experiences may either be *scheduled*—where elements are presented at specific points in the programme, *partially-scheduled*—where groups or chapters of content are delivered at points in the programme, but a user may peruse these at their own pace, and *unscheduled*—where the user has complete control over the pace of the content. The amount of interaction from the user is highlighted as an important factor by which these experiences may vary but is not classified within the taxonomy. The amount of interaction

is viewed as a continuous scale on which, at one end, a user is completely passive and is displayed additional information, whereas, at the other end, interaction is required to push the experience forward.

To contextualise some of these distinctions, the ‘Autumnwatch’ companion (Jones, 2011) provides a useful example. This package of content comprises images, diagrams and short passages of text. This content has been selected due to its relevance to the main programme and has been created or curated to accompany it. It is, therefore, companion content. Jones (2011) referred to users using this during the main programme and the companion content being delivered as a “linear sequence” alongside the main programme. As the experience occurs at the same time as the main programme, it is a synchronous experience. Furthermore, as the content has been put together for this express purpose, it is clearly orchestrated. The content used in this experience is—for the most part—clearly pre-defined and static, which makes it a fixed content experience. There is, however, also mention of image galleries. If these galleries were populated from a dynamic source (e.g., the highest rated viewer photographs from the website), this would be an evolving content experience. The experience is also considered scheduled, as the content was presented as a linear synchronised sequence alongside the show. In later work using the same content, Brown *et al.* (N.D.) compared the synchronised experience with one in which the users were provided with all of the content, which they could browse at their own pace. This variation is considered as an unscheduled experience. Jones (2011) refers to the content as being organised in chapters. Had the content been delivered as chapters that the user could navigate through, this would have been considered a partially-scheduled experience. Jones (2011) also makes some mention of the application being used after the user had finished viewing the main programme. As this usage is not alongside the main programme, it would be considered to be an asynchronous experience of companion content. An additional piece of terminology, not discussed within (Hoare & Hinde, 2016), is introduced within this thesis to refer to the content presented in scheduled or partially-scheduled orchestrated synchronous companion experiences that are not present in the main programme. This is referred to as secondary programme content (SPC) to separate it from general purpose companion content and represent the closer link that is likely to be present between the SPC and the content of main programme content (MPC) at the time at which it is presented.

### 2.4.3 Adaptive programming: object-based broadcasting

Within traditional broadcast models, a television programme can be viewed as a single continuous stream of content. VOD has altered this view by removing the programme from the broadcast stream and isolating it as an atom of content. Internet-based delivery introduces the possibility of splitting this atom into its most basic building blocks and allowing new, alternate versions to be created. This is the concept of object-based broadcasting (OBB) (Armstrong *et al.*, 2014). In the traditional workflow for creating a television programme, editors assemble a sequence of clips from different videos and combine layers of audio from many different sources. When finished, these separate clips and audio stems are rendered to one video stream and one audio stream that are synchronised. The idea of OBB is that, rather than delivering pre-rendered streams of audio and video, the clips and stems are delivered and rendered for each user.

While this may superficially seem to be merely a less efficient method of delivering a programme, this late rendering allows alterations to be made to a programme's edit for an individual user. With this method it may be possible to adjust structural elements of the programme to provide users with content of variable duration (Armstrong *et al.*, 2014). Furthermore, it may be possible to adjust more subtle features of the programme to suit the user, their device, or viewing context. This could take the form of modifying the grading, shot choice/timing, music choices (BBC Research & Development, n.d.c) or altering the audio mix to provide an optimised experience for users with hearing impairments (Shirley & Oldfield, 2015). There is obviously potential for such factors to be controlled using explicit interactions (i.e., the user selects appropriate options from a list). Some researchers advocate 'perceptive media', however, which is where information sensed about the user and their context is used to adapt the content accordingly (Forrester, 2012; Gradinar *et al.*, 2015).

While this project is not directly concerned with the format of programmes, this concept of adaptive programming and personalisation could have a dramatic impact on the user experience of television and offer opportunities for designing a new generation of television interfaces.

### 2.4.4 Input mechanisms

For many years, television user experiences have been inextricably linked with the remote control. The traditional remote control is a handheld device designed solely for the purpose of

controlling a television wirelessly. These devices usually provide the user with sets of buttons for typing channel numbers, navigating and selecting items from menus, and controlling basic TV set options. With the increased computing capabilities of current and next-generation television sets, however, a much larger range of interaction methods is possible.

The connection to a home network that facilitates the use of other connected devices to act as peripheral displays can also convey control messages back to the television set. The use of remote-control apps on portable or semi-portable devices is one obvious use of this capability. Some manufacturers of connected devices have already started to do this, offering downloadable remote control apps for smartphones or tablets (e.g., Matt (Samsung), 2012; LG USA, n.d.). Some researchers have highlighted the potential of using smartphones/tablets to provide remote controls that can be adapted for specific user activities (Lin *et al.*, 2012) and individual users (Bernhaupt & Pirker, 2014). It is important to note, however, that the possibilities of IP-based controls stretch beyond the smartphone and tablet. Just as internet of things (IOT) devices may be used in the display of information, they may be used for user input. In principle any connected device which the user could interact with and could have a control app installed on it is capable of providing some remote control functionality. Alternatively, some manufacturers have looked towards including more functionality within the television itself using gestural or speech commands (e.g., Samsung, n.d.).

With portable and semi-portable device-based television experiences, the user is unlikely to require a remote control, due to the short viewing distances. This, and the variable locations of use, make network-based control less important with these devices. As many of these devices already have touch screens, cameras and speech recognition functionality, however, there are still a host of input mechanisms that may be used.

## 2.5 Discussion

This chapter presents an overview of the current state-of-the-art in television broadcast, devices, interfaces, experiences and input controls. Television is clearly in the process of experiencing a major shift away from mass delivery of traditional linear broadcast streams to delivery over IP. While this shift may have impacts on many elements of television production, its impact on the television user experience is the focus of interest for this thesis.

### 2.5.1 Programme delivery

Section 2.2 introduces the current state-of-the-art in television programme delivery. Though traditional digital broadcast technologies (i.e., terrestrial, cable and satellite) are still important parts of the television infrastructure at the time of writing, internet delivery offers experiential benefits due to its compatibility with a range of devices, and the addition of a return channel. It seems that internet delivery is set to be an increasingly important element of television in the near future. As the delivery over IP removes most of the restrictions associated with broadcast (e.g., bandwidth), this thesis will not concern itself with how to deliver the component parts of the proposed displays to end-user clients.

### 2.5.2 Devices

With the move towards internet broadcast, the range of devices that may be associated with television experiences has expanded considerably. Section 2.3 discusses the range of technological limitations, use-contexts and user-behaviours associated with the various classes of device. While these differences raise many interesting questions, there are simply too many to be considered within the scope of a single project. It is therefore necessary to define some limitations on which factors will be considered. Within this project, the main focus will be on connected televisions, though consideration will also be given to the ramifications for display on other devices. There are several motivations for choosing to focus on the connected television use case rather than those of the portable or semi-portable devices. Firstly, at the time of writing, most homes in the UK still have a television set (Noland & Truong, 2015) and most watching of video content still occurs on either traditional or connected televisions (Ofcom, 2016). Secondly, the connected television represents the ‘ideal’ television viewing scenario. From a technical standpoint, these devices are designed for the primary purpose of displaying video and, therefore, it is assumed that the video and audio quality is reasonable. Unlike the variable use-contexts of portable and semi-portable devices, within the living room scenario it is likely that any distractions from the television are self-motivated. From an engineering perspective, it seems sensible to start with this ideal context before applying the learning to more complex scenarios.

### 2.5.3 Input mechanisms

Traditional remote controls and more recently proposed alternatives are discussed in Section 2.4.4. While the focus of this project is on the display of information rather than on user input, it is undeniable that there are cases where the nature of the interaction makes it impossible to separate the two. With the move towards a connected home, there appears to be enormous scope for control mechanisms for connected televisions besides the traditional model of a remote control. Furthermore, as this project is attempting to focus on the future of television experiences, the consideration of input devices with the proposed systems will not be limited to those in current use for interfacing with television systems.

### 2.5.4 Navigation

Interfaces for the navigation of options within television menus are discussed in Section 2.4.1. Traditionally, menu interfaces have played an important part in navigation options, both within computing and television. The need to navigate large numbers of options appears likely to be an increasing problem as television experiences expand in terms of the amount of content provided and the options available. Providing search functions is one important step that can be taken to allow users to find desired items more quickly. The success or failure of these search interfaces is likely to depend on the manner in which search requests are input and the intelligence of the search algorithm that is implemented. As this project is focused on the display of information, the design of improved search algorithms is considered out of scope. These functions, however, are unlikely to remove the need for menu representation in television interfaces due to poorly defined search criteria and users wishing to browse the available options.

Three types of menu are described which differ in terms of the types of content they present and the level of use they are likely to see. Without non-visual access to these menus, users who are unable to visually attend to the screen will find it difficult to access important settings and choose desired services or programmes. Programme and service selection menus both have important impacts on users' experiences of television systems as a result of the high frequency of their use. Due to the importance of these interface components in current and future television systems, the non-visual design of menus is explored further within this thesis.

### 2.5.5 Additional media activities

The growth of additional media activities within television user experiences raises a number of interesting questions for both visual and non-visual user interface designers. As highlighted in Section 2.4.2, a huge variety of different content and use contexts exist within this broad categorisation. Experiences comprising unrelated content are difficult to consider from a design perspective, as they could comprise any number of separate activities. Furthermore, even in synchronous use it is difficult to determine what a desirable display should enable, as users are likely to have very different goals and may be engaging with each element of the experience (i.e., the programme and the other content) to differing degrees. Asynchronous experiences are likely to take the form of websites and applications. It seems likely that more general research projects focussed on application and browser accessibility will be better placed to offer improvements to the non-visual experiences of these elements. These experiences are, therefore, also not considered within this thesis.

Related synchronous experiences are an interesting use case to consider from a non-visual perspective. Content which is experienced alongside the main programme raises the question of how information can be presented without disrupting the user experiences. Within this use case, even participants who are visually attending to the screen displaying the main programme may benefit from the presentation of the additional content in a different modality due to restrictions in the available screen space and the limitation of only being able to look at one thing at a time. Experiences in which the content has been specially curated or created, and timed to fit with specific parts of the show (scheduled fixed orchestrated experiences) would appear to be the most easily studied of these experiences. As the experience can be tightly controlled by the designer, the impact of different display elements can be more closely studied. The display principles involved for this type of experience will have implications for the display of other related synchronous experiences. This project, therefore, further explores how these experiences may be facilitated non-visually.

This thesis concentrates on the design of auditory displays for menus (Chapter 5) and scheduled orchestrated companion experiences (Chapters 6 and 7). In order to contextualise the design of the auditory displays, the next two chapters provide background on the human auditory system and perceptual considerations (Chapter 3) and design methodologies for auditory display (Chapter 4).





# Chapter 3

## Audition

### 3.1 Introduction

Sound is the compression and rarefaction of particles in a medium. Through these vibrations, we are able to gain an understanding of our environment and communicate with others. The conversion of pressure fluctuations at the ear to our perception of sources is facilitated by the human auditory system, which comprises both physical and psychological elements. As sophisticated as these systems are, they have implicit limitations. We can, for instance, only perceive sounds within a restricted range of frequencies ( $\approx 20 - 20,000$  Hertz (Hz) (Plack, 2014)) and pressure levels (120 decibels (dB)), above which there is an onset of pain and risk of hearing damage (Plack, 2014). It is therefore important that designers of auditory displays consider the perceptual limitations of human audition.

Speech is an important signal for consideration within auditory displays and represents a special case within human auditory perception. As a type of sound that is commonly used to convey information between people, it is particularly interesting from an auditory display perspective.

This chapter provides an introduction into the important aspects of human auditory perception of general sounds before focussing specifically on speech perception. Particular attention is paid to factors affecting perception of speech in acoustic scenes comprising multiple talkers.

## 3.2 Hearing: key principles

While it may be possible to embed a huge amount of information within an audio signal, if a listener is unable to hear the sound or interpret its complexities, it serves very little purpose. An understanding of sound and human auditory perception is, therefore, key to the successful design of auditory displays. This section provides an introduction to human auditory perception for the purposes of auditory display design. An introduction to the anatomy and physiology of the auditory system is provided before moving onto important perceptual considerations including: spatial perception, masking, auditory streaming, attention, and cross-modal effects.

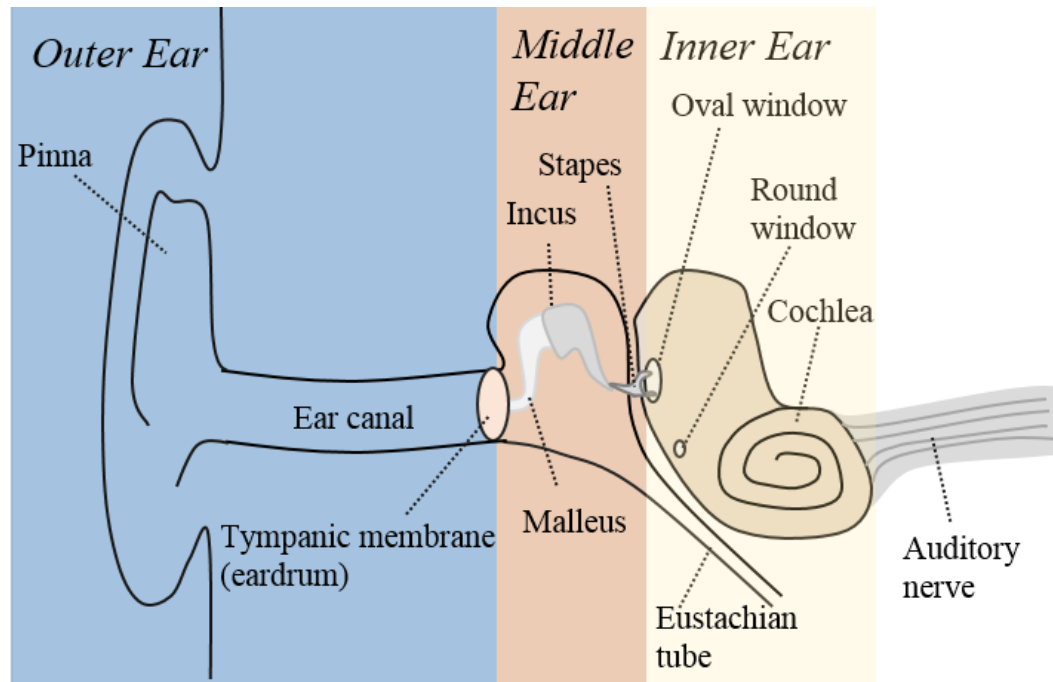
### 3.2.1 Physiology of the human auditory system

The auditory system comprises both the periphery, which deals with transduction of acoustic signals into a neural signal, and the auditory pathway, which is responsible for the processing of this neural signal (Pulkki & Karjalainen, 2015). The periphery consists of three individual segments: the outer ear, middle ear and inner ear (Plack, 2014) (see Figure 3.1). Each segment contributes towards the conversion of compressions and rarefactions in the surrounding medium into the neural representation that governs perception of sounds in space. This section provides a brief overview of how sound is processed as it travels through the auditory system.

#### The outer ear

As sound arrives at the ear, the first structure that it meets is the *pinna*, which is the external part of the ear. The pinna's complex structure makes slight modifications to the spectral characteristics of sounds depending on the direction from which the sound reaches the ear (Plack, 2014). These spectral modifications are decoded within the auditory pathway to determine the location of sound sources (Plack, 2014).

From the pinna, sounds enter the *ear canal* (*external auditory meatus*), which is a short tube which has an opening in the pinna at one end and is terminated by the *eardrum* (*tympanic membrane*) at the other end (Plack, 2014).



**Figure 3.1:** Annotated diagram of the human ear indicating the outer, middle, and inner ear sections. Adapted from (Plack, 2014, p. 54)

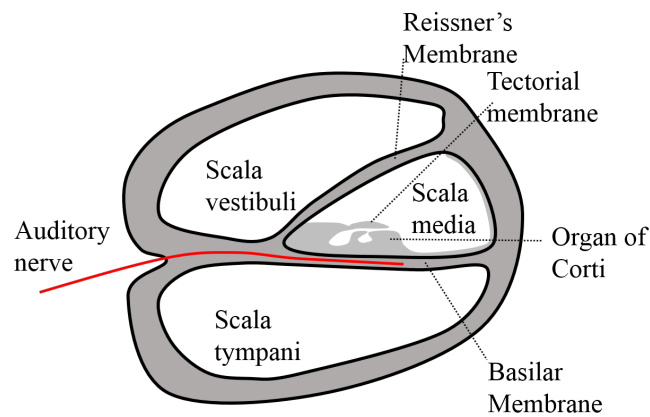
### The middle ear

The pressure variations in the ear canal cause the eardrum to move, converting the sound waves into mechanical vibrations (Plack, 2014). The eardrum is then connected to a series of small bones: *incus*, *malleus*, and *stapes*, collectively referred to as the *ossicles* (Plack, 2014) (shown in Figure 3.1). The incus is attached to the eardrum and conducts vibrations on to the malleus which passes them on to the stapes. The stapes connects the middle ear to the inner ear and is attached to the oval window at the base of the *cochlea*. The middle ear serves as an impedance mechanism to improve the transfer of energy from the air to the inner ear (Pulkki & Karjalainen, 2015).

### The inner ear

The inner ear comprises the *cochlea* and the *semicircular canals* (Pulkki & Karjalainen, 2015). As the latter is not involved in hearing (Pulkki & Karjalainen, 2015), this section focusses on the function of the cochlea.

The cochlea is a spiral structure, formed from a coiled fluid-filled tube (Plack, 2014). The tube itself is constructed of three channels, the *scala vestibuli*, the *scala tympani* and the *scala media* (Plack, 2014) (see Figure 3.2). The oval window is at the base of the *scala vestibuli*.



**Figure 3.2:** Annotated cross-section of the cochlea. Adapted from (Plack, 2014, p. 56)

The scala vestibuli is separated from the scala media by Reissner's membrane and the scala media and the scala tympani are separated by the *basilar membrane* (Plack, 2014). The scala vestibula and the scala tympani contain the same fluid and are joined by a hole at the top of the spiral (Plack, 2014).

At the base of the scala tympani there is a membrane, similar to the oval window, called the round window (Plack, 2014). Vibrations of the oval window cause compression and rarefaction in the fluid within the cochlea, which causes the Reissner's membrane and the basilar membrane to vibrate (Plack, 2014). As the basilar membrane's thickness changes from the base to the apex of the cochlea, becoming wider and looser as it approaches the apex, different frequencies of vibration lead to different locations of the basilar membrane resonating (Plack, 2014). This means that sounds are broken into their composite frequencies that are represented at different locations along the basilar membrane. For this reason, it is common to think of the basilar membrane as a bank of band-pass filters.

The *organ of Corti* sits on top of the basilar membrane and comprises sets of inner and outer *hair cells* which have *stereocilia* that extend above the organ of Corti (Pulkki & Karjalainen, 2015). With the outer hair cells, the stereocilia embed into the *tectorial* membrane above, while the stereocilia of the inner hair cells do not (Plack, 2014). When the basilar membrane vibrates, so does the tectorial membrane. This movement causes the stereocilia to bend, which causes the hair cell to release a neurotransmitter (Plack, 2014). This is then received by receptors of neurons of the *auditory nerve* and causes electrical spikes (action potentials), which form the neural signal within the auditory pathway (Plack, 2014).

As neurons attach to hair cells at specific points on the basilar membrane, each neuron is associated with a specific frequency depending on the hair cell's location. In practice, this

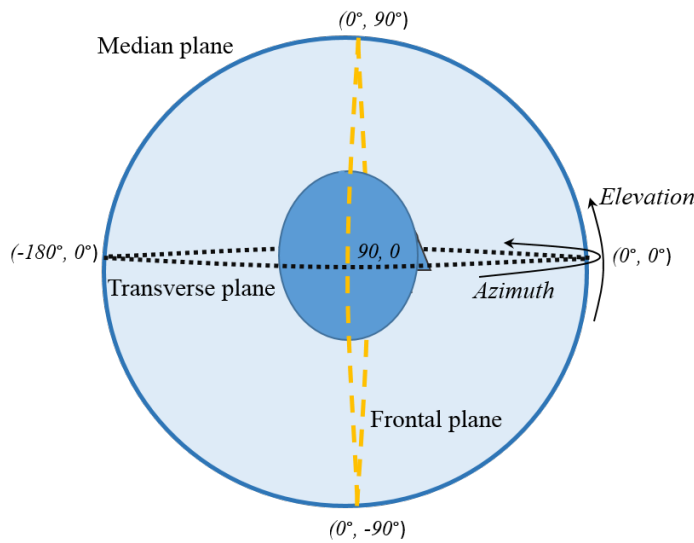
means that when a region of the basilar membrane is excited by its characteristic frequency, there is an increase in the rate of action potentials in the neurons from the hair cells at that location. The rate of the action potentials is dependent on the amplitude of the basilar membranes movement, increasing with the amount of displacement until it reaches the limits at which action potentials can be generated (Plack, 2014). In addition to conveying information about the frequency that is excited, this neural code also provides information regarding the phase vibration (Plack, 2014).

### **Auditory Pathway**

The auditory nerve runs from the cochlea to brain where the signal is decoded to inform our perception of sound. Different areas of the brain are thought to be responsible for extracting different information from the neural representation produced by the cochlea. The exact function of the different nuclei through which the signals pass, are still the matter of research (Plack, 2014). As the neuropsychology is not the primary concern of this thesis, this section does not go into the details of the each of the stages involved in this process but provides a broad overview.

Up to this point, the description has considered only the activity in a single ear. At this point it is, however, important to consider neural codes arriving from both cochlea. For each stage in the processing of the neural code, there is a pair of specialised nuclei—one on each side of the brainstem (Plack, 2014). As the neural signals are passed between specialised nuclei, important information pertaining to the location of the sound and its characteristics are extracted (Plack, 2014). After this, the resulting signals are passed to the *auditory cortex* (Plack, 2014). The auditory cortex is thought to perform the higher level processing of sounds and communicates with other parts of the brain to integrate information from the various modalities and to analyse semantics of auditory information (Plack, 2014).

This discussion has considered only afferent (periphery to cortex) signals. There are, however, also efferent connections (originating in the brain), which are thought to control the movement of the basilar membrane under high sound pressure levels, and alter the processing of sounds in the brainstem (Plack, 2014).



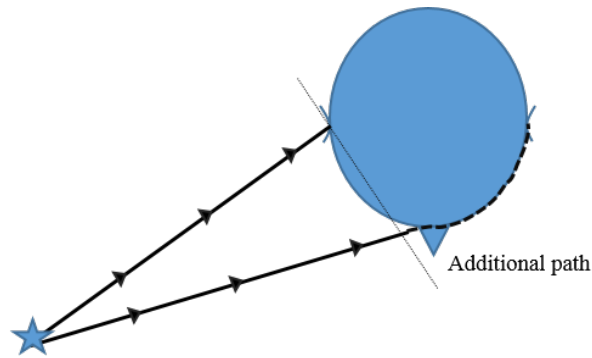
**Figure 3.3:** Diagram showing the median, transverse and frontal planes. Adapted from (Blauert, 1997, p. 14). Locations are referred to in (azimuth, elevation) format.

### 3.2.2 Spatial auditory perception

In natural acoustic environments, sounds arrive at the ears from many different directions. Birds sing in the trees above us, our footsteps come from beneath us, and a twig snaps behind us. Being able to identify the location of a sound source is an important feature of human auditory perception. Unlike vision, audition is sensitive to sounds coming from all directions. By decoding spatial information from sounds, we are able to develop a better understanding of our environment. Furthermore, spatial information of auditory events provides information to allow quick orientation of visual attention to salient sounds. It also provides additional benefits where the perception of multiple sounds are concerned, which will be discussed later within this chapter.

Looking at a sound signal, there are no implicit cues to its location and as sounds arrive at the ears from many directions. This section describes how spatial information is deduced from the signals arriving at the ears and how headphone presentations can produce different spatial percepts.

When talking about locations in relation to a listener, it is useful to define several planes relative to the head. The *median* plane bisects the head down the centre between the eyes. The *frontal* plane also runs downwards through the head but perpendicular to the median plane, while the *horizontal* plane runs laterally through the head (Blauert, 1997) (see Figure 3.3). It is common to refer to positions in terms of their azimuth or elevation (Blauert, 1997). Azimuth refers to their angular displacement on the horizontal plane with  $0^\circ$  directly in front



**Figure 3.4:** Illustration of the additional path taken by sound to reach the contralateral ear for sources away from the median plane. Adapted from (Rumsey, 2001, p. 22)

of the listener on the median plane, and  $180^\circ$  directly behind on the median plane. Elevation refers to angular displacement along the vertical, median plane, with  $0^\circ$  at the level of the ears,  $+90^\circ$  directly above the head and  $-90^\circ$  directly below.

### Spatial cues

As mentioned in Section 3.2.1, the outer ear imparts spectral modifications on signals depending on the angle from which they enter. This spectral modification occurs because sound takes different routes around the pinna before entering the ear canal. As these routes differ in length, phase differences are introduced between the signals that have taken different paths. When these signals are combined, the phase differences result in attenuating some frequencies, while accentuating others (Moore, 2012). Due to the size of the pinna, this only effects higher frequencies, above 4 kilohertz (kHz) (Plack, 2014),

This, along with reflections from the head and shoulder, create monaural cues that are highly individual, due to anthropometric differences. These cues are particularly important in resolving the elevation of sources (Plack, 2014). Studies in which listeners have attempted to localise sound sources with modified monaural cues have shown an increase in the amount of error (Hofman *et al.*, 1998). It is interesting to note, however, users can adapt to these modifications that over longer periods (Hofman *et al.*, 1998).

In addition to monaural cues, spatial hearing also exploits small differences between the signals arriving at both ears. There are two principle *binaural* cues that vary depending on the location of a sound source. The first of these is differences in the time at which the signal arrives each ear, referred to as the interaural time difference (ITD) or interaural phase difference (IPD). When sounds are away from the median plane, the source will be closer to one of the ears (see Figure 3.4). As a result of this, the sound will reach the closer

(ipsilateral/proximal) ear first. The magnitude of the ITD increases with angular separation from the median plane, reaching a maximum of approximately 690 microseconds (Moore, 2012). At low frequencies, this is exhibited as phase difference between the signals due to the period of the wave (Moore, 2012).

The second of the binaural cues is the difference between the levels of the signals arriving at each ear, the interaural level difference (ILD). The ILD is largely due to shadowing caused by the head, which causes a reduction in level at the further (contralateral/distal) ear (Moore, 2012). The amount of ILD is dependent on the frequency of the sound, with higher frequencies showing larger differences than lower frequencies (Feddersen *et al.*, 1957). Unlike ITD, ILD shows a more complex relationship to azimuthal distance from the median plane, which depends on their frequency (Feddersen *et al.*, 1957).

Rayleigh (1907) noted that the two cues are effective for only certain frequency ranges. At higher frequencies ( $> 1500$  Hz), the wavelength of signals are shorter than the distance between the ears, making the phase differences unreliable (Moore, 2012). Similarly, at low frequencies the wavelength of the signals means that the signal diffract around the head, making the level difference unreliable (Moore, 2012). It would seem, therefore, that neither cue covers the entire frequency range. At high frequencies, however, time differences in the envelope of the signals can be observed (Middlebrooks & Green, 1990). These have been found to provide localisation cues (e.g., McFadden & Pasanen, 1976; Henning, 1980) and, therefore, may work alongside the ILD at higher frequencies.

### **Simulating spatial sound with headphones**

When presenting a signal directly to each ear using headphones, binaural and monaural cues are not present and sources are perceived as coming from within the listener's head (Plack, 2014). When the signal is identical at both ears, this is referred to as a *diotic* presentation. In order to provide users with an accurate spatial impression of sounds presented over headphones, it is necessary to introduce the monaural and binaural cues that would be present normally.

As the acoustic cues that influence auditory spatial perception for a source emanating from a specific location are attributable to the linear time invariant response of the head, shoulders and pinnae, it is possible to accurately model the system through the use of impulse responses (Blauert, 1997). These impulse responses are referred to as head-related impulse responses (HRIRs) in the time domain or head-related transfer functions (HRTFs) in the frequency



domain. Through convolving a source audio signal with the HRIR and then presenting the resulting signals over stereo headphones very convincing spatialisation can be achieved. Within this thesis, this is referred to as binaural presentation, processing or rendering. The results of binaural rendering are, however, highly variable and depend heavily upon both the methods used to capture the HRTFs and to render the final acoustic scene. Factors such as the individual characteristics of the HRTF and whether head-tracking systems are used to recreate cues from small head movements are known to effect the quality of the spatialisation (Rumsey, 2001).

Binaural processing is not the only method for providing some spatial impression over headphones. By including only differences in intensity or phase between the ears, it is possible to shift the perceived location of the source from the centre of the head. The manipulation of the intensity at each ear causes the signal to remain within the head but shift towards the ear presented with the higher intensity signal (Blauert, 1997). This is referred to as intensity panning, and is commonly used in consumer systems. Stereo mixes rely on differences in intensity (Moore, 2012). While presentation over properly configured loudspeakers leads to phase differences being reconstructed due to cross-talk between the channels (Blumlein, 1933), this is not the case when stereo material is listened to over headphones.

An extreme form of this is *monaural* presentation, where the signal is only presented to one ear, or *dichotic* presentations, where completely different signals are presented at each ear. While these are rarely encountered in consumer systems, they have formed the basis for experimental work on spatial auditory attention.

### 3.2.3 Masking

Masking is when the “sensitivity for one sound is affected by the presence of another sound” (Gelfand, 2010, p. 187). There are several different forms of masking that can occur depending on the timing and frequency of the sounds and required listening task. A broad distinction is drawn between *energetic* and *informational* masking, though some take issue with this terminology (cf. Moore, 2012).

Energetic masking is defined as interference between two signals that is attributable to the physiological limitations of the auditory system prior to the auditory nerve (Durlach *et al.*, 2005). Energetic masking, therefore, refers to when the action potentials in the neurons from the cochlea do not differ sufficiently from those produced by the masker in isolation when the target signal is added (Moore, 2012).

Simultaneous energetic masking usually occurs when target and masker signals have similar frequencies. Even a pure tone excites areas of the basilar membrane surrounding the part with the corresponding characteristic frequency (Moore, 2012). Simultaneous masking is thought to be caused when the masker causes sufficient residual excitation at the location of the target to obscure its presence (Moore, 2012). Put another way, thinking of the basilar membrane as a bank of band-pass filters, this masking occurs when the energy from the masking signal within the bands where the target is present is high enough to obfuscate it. The bandwidth of these filters and their shape affects how close signals can be before masking occurs. These filters are commonly referred to in terms of their equivalent rectangular band (ERB), which increases with the centre frequency of the filter. Glasberg & Moore (1990) developed an equation to calculate the ERB for a given centre frequency ( $f_c$  in kHz):

$$ERB = 24.7(4.37f_c + 1) \quad (3.1)$$

There is also some evidence that additional masking can occur due to activity further from the target location on the basilar membrane, due to the masker suppressing the neural activation from the target signal (e.g., Sachs & Kiang, 1968; Delgutte, 1990; Moore & Vickers, 1997).

Thus far, masking has been considered only in a monaural context. As spatial sound perception demonstrates, the differences between signals arriving at both ears can provide useful additional cues. Binaural differences can play an important role in reducing the effects of masking. Presenting a target signal with a different phase in each ear, or removing the target signal from one channel, results in less masking from diotic noise than when the target was also presented diotically (Moore, 2012).

Informational masking is less well understood than energetic masking and its precise definition has been the matter of some debate (Yost, 2006; Micheyl, 2006; Watson, 2006). The loosest definition of informational masking is that it refers to any masking that cannot be explained by energetic masking (Micheyl, 2006). This suggests that it occurs after the transduction of signals to neural code by the cochlea. Although the precise cause of informational masking is not clear, it is attributed to a failure of selective attention either due to distraction by masking signals or a failure to segregate the streams (Micheyl, 2006). This definition of informational masking as the inability to segregate due to a streaming failure is similar to what Bregman (1990) referred to as *fusion*. The difficulty in defining informational masking is likely to be, in part, due to the likelihood that several different processes are involved (Moore, 2012).

Informational masking is commonly referred to when discussing concurrent speech. It is, however, also a factor with non-speech stimuli (Neff & Green, 1987; Neff *et al.*, 1993; Durlach *et al.*, 2005). Release from informational masking has been found from reducing uncertainty in the task or the *similarity* of competing non-speech sounds (Kidd Jr. *et al.*, 2002; Arbogast & Kidd, 2000).

One may also distinguish between simultaneous masking, where both target and masker occur at the same time, and non-simultaneous masking, where the masker occurs at a different time to the target signal. With non-simultaneous masking, the perception of the target is affected by a masker that either precedes (*forward masking*) or follows it (*backward masking*). The effects of forward masking can persist for up to 200 ms and, similarly to simultaneous energetic masking, it is dependent on the similarity of target and masker frequencies (Moore, 2012).

The backward masking of detection tasks is limited to instances in which a masking signal occurs within approximately 20 milliseconds (ms) of a target (Oxenham & Moore, 1994). The effects on recognition tasks (backward recognition masking (BRM)), however, have been reported to be considerably longer, with optimal performance usually reported as around 200-350 ms (Massaro, 1970, 1974; Sparks, 1976; Massaro & Idson, 1977). Interestingly, the performance in these tasks has been found to be improved by musical experience (Sparks, 1976), and has been observed when target and masker are presented dichotically (i.e., each is presented to only one ear) (Massaro, 1970)—which demonstrates that it occurs after generation of neural code by the cochlea.

#### 3.2.4 Auditory streams

As discussed in the previous section, many questions remain unanswered about how the human brain decodes information from the neural representations generated in the cochlea. From studies on human experiences of sound, however, it is possible to identify important perceptual features. When we listen to our environment, we don't think of phase and spectrum, but sources in space. Consider a rattle that is shook once, each individual bead impacts with the internal surface of the rattle creating impulsive sonic events. The sound waves then travel through the air via various routes, some travelling straight to the ear, others bouncing around the room first. This means that the ear would receive many individual bead impacts and their numerous reflections. This is, however, perceived as only one auditory event—a single shake of a rattle. Similarly, if someone were to repeatedly tap on a wood block, though we would be able to tell there were discrete percussive events, we would group

them together over time as a single rhythmic sequence.

Bregman & Campbell (1971) originally introduced the concept of *auditory stream segregation* and used them to explain the reduction in performance at judging the order of short repeating tones sequences when the tones' frequencies varied by larger intervals compared to when they were similar. Bregman & Campbell (1971) suggested that the tones split into distinct streams. In his book on the subject, Bregman refers to an auditory *stream* as "the perceptual unit that represents a single happening" (Bregman, 1990, p.10). Using this terminology, the shake of the rattle would be one perceptual stream, as would the tapping of the wood block. If a duck quacked between the taps of the woodblock, this would be perceived as a separate stream. It is obvious on a conceptual level that these are different sources. Considering the information available from the acoustic cues alone, it is important to consider what causes the woodblock and the duck to be grouped into separate streams. Bregman (1990) drew inspiration from Gestalt psychology (e.g., Koffka, 1922) and put forward the idea that evidence is drawn from a collection of cues, which informs the decision about whether an individual element belongs to a specific stream. Within visual perception this can mean that distinct objects provide the impression of a single object. The cues used for grouping sounds are, of course, different to those found within visual perception. In his comprehensive review of the research on the perception of auditory streams, Bregman (1990) differentiates between *sequential* integration, which describes the combining of elements into a stream over time (e.g., the rhythmic tapping of a woodblock), and *simultaneous* integration, which describes the grouping sounds that occur at the same time (e.g., the components of complex tone).

Sequential integration is largely affected by the similarity of the sounds to what came before (Bregman, 1990). Similarity, however, can be defined by a number of distinct acoustic cues. Bregman (1990) identifies the proximity of pitch, timing, intensity, timbre, spatial location, and consistency with the apparent *motion* of the stream up to that point as important factors in deciding the organisation of perceptual streams. These cues, however, do not exist in isolation but may compete or collaborate with each other for different perceptual groupings (Bregman, 1990). Generally, the larger the differences between successive auditory events the greater the likelihood of them being perceived as two separate streams. In the woodblock-duck scenario, the duck's quack differs in terms of several of these factors causing it to be perceptually segregated. Bregman (1990) notes that this formation of streams affects a listener's ability to judge relationships between constituent sounds, particularly with rhythm and ordering.

Simultaneous integration is concerned with grouping energy at individual frequencies (*partials*) into more complex sounds. Bregman (1990) notes that shared onset times, perceived locations, the difference in frequency, as well as amplitude and frequency trajectories were important for grouping partials. Furthermore, where partials were harmonically related to a shared fundamental frequency, they were more likely to be grouped (Bregman, 1990). In addition to this, he refers to the “old-plus-new” heuristic where a partial of a complex tone is attributed to a preceding sound (Bregman, 1990).

In addition to these classes of segregation, Bregman (1990) also differentiates between *primitive* organisation and *schema-based* organisation. Primitive organisation is based on universal acoustic laws (e.g., a sound is unlikely to change suddenly) and assumed to be innate, whereas schema-based organisation is based on the experience of the listener (e.g., the language they speak) or the listener actively attempting to attend to something. Bregman (1990) characterises the difference between the two as primitive organisation partitioning the auditory information, which is contrasted against schema-based organisation selecting elements from the mixture.

It should be clear that auditory stream theory is an important consideration in the design of auditory display. A poorly-considered auditory display may lead to elements that were supposed to appear as separate fusing, or a single element splitting into separate perceptual streams. Either case would be extremely problematic for users of the system.

### 3.2.5 Auditory attention

In the previous sections on masking and auditory stream formation, there has been some discussion of the effects of attention. Attention is an important factor in considering perception, as it is the process that chooses what information is selected to be processed further and what information should be discarded (Smith & Kosslyn, 2014). The consideration of auditory attention is clearly important for the design of auditory displays. It is noted, however, that this is a large field of study that still consists of many unknowns. This thesis is, therefore, only able to provide a brief overview of some of the important considerations of auditory attention.

When confronted with multiple concurrent sources, users can decide to attend to one source, referred to as *selective* attention, or split their attention between multiple sources (*divided* attention) (Styles, 2006). A famous example of selective attention was introduced by Cherry (1953) as the “cocktail party problem”, where a listener attends to the speech from one talker

in the presence of speech from other talkers. In the divided attention equivalent, the listener would attempt to attend to two talkers simultaneously (Styles, 2006).

How attention is oriented may be decided due to different factors. In some instances a listener may consciously choose to devote attentional resources to a specific item, referred to as an *endogenous* shift (e.g., a conversations they are interested in), or a salient event may draw attention towards it, referred to as an *exogenous* shift (e.g., a balloon popping) (Posner, 1980).

Auditory attention can be oriented to specific frequency ranges. When participants are given a priming tone and asked to detect a probe tone in noise, participants exhibit decreasing performance as the difference between the pitch of the prime and the probe increases until detection rates are interpretable as chance (Greenberg & Larkin, 1968; Scharf *et al.*, 1987; Dai *et al.*, 1991; Botte, 1995). Similarly, participants exhibit decreasing performance in indicating probe durations as their frequency deviates from the prime (Mondor & Bregman, 1994). It would appear that the width of these *attentional bands* (Dai *et al.*, 1991) can alter depending on the bandwidth and level of the priming tone (Botte, 1995). Others have found results that suggest participants can monitor multiple frequency regions at once (e.g., Schlauch & Hafter, 1991).

Given the auditory system's ability to interpret spatial information, it is also feasible that listeners may be able to orient attention to specific locations. A number of research studies appear to suggest that auditory attention can be oriented to locations by auditory cues (e.g., Ward, 1994; Mondor & Zatorre, 1995; Spence & Driver, 1994, 1997). Furthermore, accurate information about the location of the target has been found to improve response time and accuracy for identifying non-speech patterns when multiple concurrent stimuli are present (Arbogast & Kidd, 2000).

### 3.2.6 The influence of vision

Audition does not exist in isolation from the other senses. In Section 3.2.1, it was mentioned that signals from the auditory cortex were sent to different areas of the brain for further processing with information from other senses. This can mean that the experience of isolated audio changes when visual information is incorporated.

The perception of the location of a sound source can be altered by the presence of visual stimuli (e.g., Wallach, 1940). A famous instance of this is where the presence of a synchronized

visual event occurring at a different location causes the perceived location of the sound to move towards it, referred to as the *ventriloquism effect* (Howard & Templeton, 1966). This can be observed when non-speech stimuli are presented alongside other synchronised visual stimuli (Bertelson *et al.*, 2000; Alais & Burr, 2004). It is notable that where visual location is ambiguous, auditory stimuli can also influence the perceived location of the visual stimulus (Alais & Burr, 2004).

Bregman (1990) noted that, alongside the cues discussed in Section 3.2.4, visual cues could also affect sequential integration. As highlighted by Bregman (1990), O’Leary & Rhodes (1984) demonstrated this in a study which found that a sequence of tones was perceived as two streams when presented with a concurrent visual stimuli that appeared as two objects, but when the visual stimuli was changed so as to appear as a single object the auditory streams combined.

Numerous studies have demonstrated links between auditory and visual spatial attention (e.g., Gopher, 1973; Reisberg *et al.*, 1981; Driver & Spence, 1994; Spence & Driver, 1996, 1997; Rorden & Driver, 1999; Spence *et al.*, 2000; Blurton *et al.*, 2015). It is common for us to look at what we are listening to. In fact, Gopher (1973) noted that attending to a spatial auditory source can influence eye movements. He observed that participants’ eye movements changed when they were attending to temporally-distinct dichotic speech, making fewer spontaneous eye movements but more voluntary eye movements towards the source (Gopher, 1973). Rorden & Driver (1999) found that participants exhibit faster judgements of sound elevations when they move their eyes towards the side of the sound. This suggests that listeners tend to orient visual and auditory attention in the same direction.

Others have also found facilitatory effects of visual and auditory attention being oriented in the same direction. Spence & Driver (1996, 1997) performed experiments on orienting visual and auditory attention without altering gaze, referred to as *covert* attention. Spence & Driver (1996) performed an experiment in which participants had to judge the elevation of a light or sound at a lateral position when presented while their gaze was fixed at a central point. It was found that providing cues about the side that the task would be on improved performance even when the modality of the task was different to what they were led to believe. This was taken to indicate a link between the auditory and visual spatial attention. It was, however, also found that visual and auditory spatial attention could be split when participants expected tasks of different modality on each side (Spence & Driver, 1996). A following study by Spence & Driver (1997), found that in the same task, uninformative auditory cues led to

better performance when presented from the same side as the visual task, though visual cues were not found to facilitate auditory judgements. More recent work has found that exogenous visual cues can draw auditory spatial attention, but to be effective the cue and target must be closer together than when auditory cues are used (Prime *et al.*, 2008).

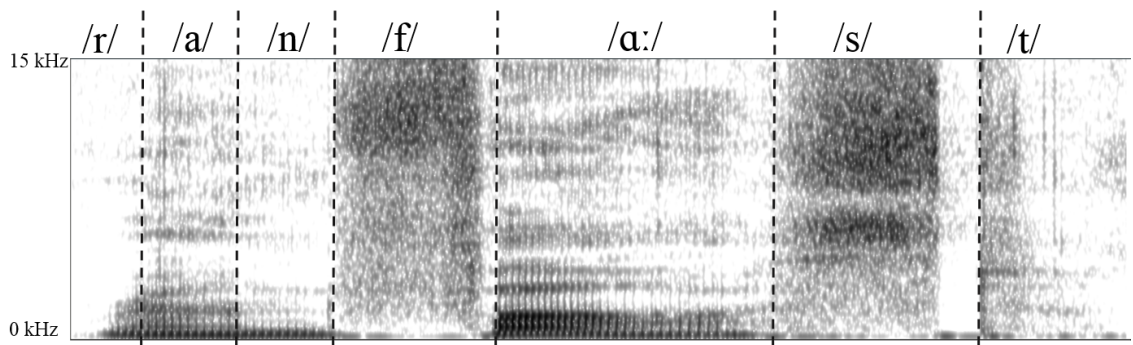
### 3.3 Speech perception

As social creatures, speech plays a vital role in our ability to understand others and be understood. Its importance is such that the first stages of language learning begin before birth (Moon *et al.*, 2013). Furthermore, the role of speech in human communication makes it interesting from the perspective of auditory display. This section, therefore, provides a short introduction to speech signals and how they are understood by human listeners. Multiple talker scenarios are also considered, and important factors that affect speech perception in these scenarios are discussed.

#### 3.3.1 Speech as a signal

Speech is traditionally split into phonemes, which are the smallest units that are present in languages that can be used to contrast between words (Moore, 2012). This thesis uses the International Phonetic Alphabet (IPA) standard for representing phonemes, where symbols are enclosed by a slashes. These phonemes may be voiced or unvoiced. The voiced sounds are created when air is forced through the vocal folds which creates a periodic signal over a wide frequency range (Moore, 2012). Through manipulating the shape of the vocal tract, talkers alter the spectral characteristics of this signal to create different voiced speech sounds (Moore, 2012). Voiced sounds, therefore, comprise a fundamental frequency ( $F_0$ ) and a series of spectral peaks, referred to as formants (Moore, 2012). Voiceless signals, in contrast do not involve vibration of the vocal folds. Speech is formed of vowels, which are generally voiced, and consonants, which can be voiced or unvoiced and involve restricting airflow at specific points within the vocal tract (Moore, 2012). Voiceless consonants create sound by restricting the airflow at different points in the vocal tract to create turbulent airflow (e.g., the /s/ in sack) or the sudden release a burst of air (e.g., the /t/ in tag). These voiceless signals are not periodic and more noise-like in character but exhibit distinct spectra (Moore, 2012). Through combining different phonemes with their individual cues, talkers form spoken words and sentences (see Figure 3.5).





**Figure 3.5:** Spectrogram of speech signal containing the phrase “ran fast” showing the approximate location of the phonemes. Darker regions denote greater amounts of energy. Formants are visible within the two vowels /a/ and /ɑ:/.

Though it is convenient to refer to phonemes as having a single identity, this is not the case. When phonemes are spoken within a word, their precise characteristics depend on the phonemes that surround them, as the vocal tract has to move between positions for each phoneme (Harley, 2014). This is referred to as co-articulation (Harley, 2014). This means that the vowel, /ʌ/, in duck is different to the one that would be found in dull.

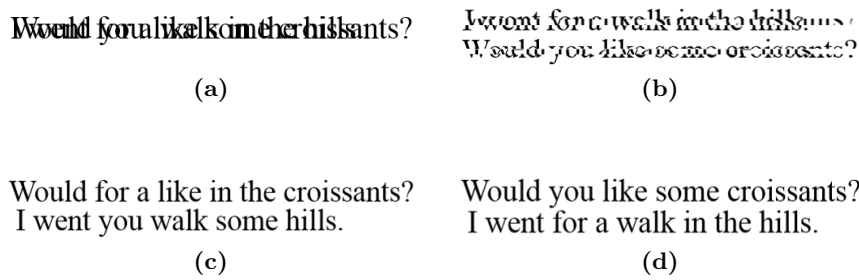
### 3.3.2 Recognising words and understanding meaning

There is some debate whether phoneme recognition needs to occur before words are recognised (discussed in (Harley, 2014)). However, in order to extract the meaning from speech, the listener must be able to identify elements of the signal which construct words.

Frauenfelder & Tyler (1987) described word recognition in different stages. In the first, the initial unit of speech signal is translated into a form which can be used to compare to the listener’s lexicon. From cross-referencing this unit with the lexicon, entries (possible words) are *activated*. As more information reaches the ear and is processed over time, one entry is identified as the best fit. The *word recognition point* then occurs when the entry is selected and the listener recognises the word.

Beyond the recognition of a word, there is also the access to the information related to it. This is referred to as *lexical access* (Frauenfelder & Tyler, 1987). Several different models of word recognition exist, which suggest that this lexical access of different information occurs at different times relative to the recognition (Frauenfelder & Tyler, 1987). An in-depth discussion of these various theories is, however, beyond the scope of this thesis. The interested reader is referred to Harley (2014, pp. 267-281).

Given the complexity of speech recognition, it is somewhat surprising to find that speech



**Figure 3.6:** Visual representations of different stream segregation failures that may occur with speech. (a) both streams are combined in one stream; (b) simultaneous streaming failure; (c) sequential streaming failure; (d) correct stream segregation

signals are very resilient to interference. This is due to the amount of redundancy in spoken communication. This occurs even at a purely acoustic level. As each phoneme is associated with a range of cues, if some are distorted or removed, other cues can still facilitate recognition (Moore, 2012). Additionally, contextual information can also influence word recognition (e.g., Warren, 1970; Meyer & Schvaneveldt, 1971; Blank & Foss, 1978; van Alphen & McQueen, 2001).

Sentence-level processing generates the syntactic information, which determines the roles of individual words and the sentence’s subject (Harley, 2014). Following this, further processing is required to *comprehend* the meaning of the sentence within its context (Harley, 2014).

### 3.3.3 Concurrent speech

A complex problem is faced in acoustic environments comprising more than one talker. Thinking in terms of stream segregation (Bregman, 1990), both sequential and simultaneous stream formation must be successful in order for speech to be understood in the presence of competing signals. This is to say that the listener must be able to assign elements of the mixture (e.g., formants and fundamental frequencies) to the correct streams of speech with sufficient accuracy to allow for successful word recognition, and also group the sequence of utterances from one talker (e.g., words, phrases) into the same stream. A streaming failure at either stage could interfere with the listener’s ability to interpret the meaning of the attended speech (see Figure 3.6).

Previous reviews on the factors affecting performance concurrent speech scenarios identified signal-to-noise ratio (SNR), target-to-masker ratio (TMR), spatial separation, number of competing talkers, and knowledge of characteristics of the target stream (Brungart & Simpson, 2002; Ericson *et al.*, 2004; Brungart & Simpson, 2005a). This discussion focusses on

the factors encountered when concurrent speech is presented without additional noise sources or manipulating the relative level of different talkers (i.e., TMR).

### **Number of concurrent talkers**

Perhaps the most obvious factor in concurrent speech perception, is the number of concurrent sources. As increasing the number of concurrent items decreases the SNR and increases the spectro-temporal density of the interference causing the probability of recovering the target phrase to be reduced (Miller, 1947), it would be expected to inhibit speech perception from purely energetic-masking. A number of studies have observed that as the number of competing talkers is increased, detrimental effects can be observed for intelligibility (Miller, 1947; Drullman & Bronkhorst, 2000), coordinate response measure (CRM) task performance (in which participants report keywords following a specified call sign) (Brungart *et al.*, 2001; Shafiro & Gygi, 2007; Nelson *et al.*, 1999), detection (Simpson *et al.*, 2006) and localisation (Simpson *et al.*, 2006; Drullman & Bronkhorst, 2000).

The number of concurrent sources which the listener is required to attend to in divided attention scenarios has a considerable impact on performance, with research finding decreasing detection sensitivity (Brock *et al.*, 2008) and identification accuracy as the number of targets increases (Shafiro & Gygi, 2007). In these situations, however, it is important to note that for every additional talker the amount of information the user is having to process is increased.

### **Pitch differences**

As discussed in Section 3.2.4, pitch is known to be an important factor in stream segregation (Bregman, 1990). Due to the different speech characteristics of individuals, it is unlikely that two voices would have identical fundamental frequencies ( $F_0$ ) in natural multi-talker scenarios. It therefore seems likely that this would be an important factor in the segregation of concurrent speech streams.

Research into the effects of  $F_0$  differences in concurrent speech perception have been generally classifiable into those focussing on vowel identification and those looking at speech intelligibility. A number of experiments based on double vowel identification have demonstrated improvements in performance with the introduction of small  $F_0$  differences of less than a semitone (Zwicker, 1984; Culling & Darwin, 1993), and asymptoting as

the difference approaches two semitones (Summerfield & Assmann, 1991; Assmann & Summerfield, 1990). However, in sentence-level tasks it has been found that performance is improved with larger  $F_0$  separations of up to three (Brokx & Nootboom, 1982), eight (Assmann, 1999) and twelve semitones (Darwin *et al.*, 2003).

The  $F_0$  in natural speech is not static but contains natural variations over time, referred to as intonation (Dobrovolsky & Katamba, 1996). In a work with concurrent vowels, it was found that time varying  $F_0$ s improved vowel identification especially when the slopes of the two  $F_0$ s were in opposite directions (Chalikia & Bregman, 1989). In Darwin *et al.* (2003), speech stimuli with natural intonation were used and found little difference between 0 and 1 semitone separations and maximal performance at a separation of 12 semitones. This contrasts with the findings of Brokx & Nootboom (1982), where the speech was resynthesised to have a constant  $F_0$ , who observed an improvement from introducing a one semitone difference and found more errors with an octave separation than when the difference was only three semitones. Darwin *et al.* (2003) attributed these differences in results to the intonation present giving transient  $F_0$  differences allowing the sources to be segregated more easily. This suggestion is partially supported by a finding that the inclusion of intonation tended to improve intelligibility with small  $F_0$  separations although significance was not achieved (Assmann, 1999). A recent experiment has also shown that the disruption of natural intonation by substituting a word from another sentence reduces performance at reporting keywords from a target sentence (Iyer & Brungart, 2010). This effect, however, appears to be limited when there is more than one competing talker (Iyer & Brungart, 2010).

### **Spatial Separation**

Spatial separation and the effects it has on the perception of concurrent speech has been one of the most rigorously researched areas concerning concurrent speech scenarios. This dates back the early work of Cherry (1953) with dichotic speech presentations. The experiments highlighted the large improvements in selective attention through separating target and competing speech streams onto different channels, when compared to diotic presentations, and the relative small amount of information from the competing stream which the listener is aware of in such scenarios. It was observed that users were only able to note acoustical phenomena from the competing stream such as changes in gender or replacement with pure tones, while factors such as language changes generally went unnoticed. This said, a later experiment within the same paper did find that users became aware of the content of the

competing speech stream when it was a delayed version of the target stream, suggesting that either semantic content or the similarity of the acoustic signal to a recent trace in memory was able to draw attention. Others have since also found cases in which some processing of the unattended stream still occurs (e.g., Moray, 1959; Treisman, 1960; Bentin *et al.*, 1995; Wood & Cowen, 1995). Following works have confirmed that spatial unmasking is present with concurrent speech streams separated on the azimuthal axis (Spieth *et al.*, 1954; Ericson & McKinley, 1997; Drullman & Bronkhorst, 2000; Ericson *et al.*, 2004; Hawley *et al.*, 1999), where both ITD and ILD provide the spatial impression. In reviews of factors affecting multi-talker displays, spatial separation has been identified as a particularly useful cue (Brungart *et al.*, 2002; Ericson *et al.*, 2004; Brungart & Simpson, 2005a).

The degree of spatial release from masking (SRM) with a spoken target is dependent on whether the masker is speech or noise (Best *et al.*, 2013; Freyman *et al.*, 1999; Noble & Perrett, 2002), the similarity of the target and masker voices (Ericson & McKinley, 1997; Helfer & Freyman, 2009; Noble & Perrett, 2002), the number of sources (Ericson & McKinley, 1997) and the listening task (i.e. detection, identification) (Nelson *et al.*, 1999). Most of these factors affect the amount of informational masking present. Studies generally show reduced SRM as the degree of informational masking posed by the competing signal is reduced. This effect was demonstrated by Best *et al.* (2013) in an experiment in which the effect of introducing a 600  $\mu$ s ITD on masking signals with a diotic speech target and the amount of spectral overlap between signals were investigated. Whilst with a Speech-Shaped Noise (SSN) masker (assumed to cause little informational masking) a small amount of SRM was observed, a speech masking signal was found to impair performance more due to the added informational masking in the diotic masker condition, but when separated achieved comparable performance to the spatial noise condition. This indicates that the addition of the ITD cue effectively removes the effect of the informational masking.

It has already been established that auditory attention can be oriented to spatial locations to facilitate processing of non-speech signals from expected locations (see discussion in 3.2.5). Experiments with spoken stimuli have also observed effects of *a priori* knowledge about the location of the sources. Information regarding the location of a target in a multi-talker mixture either through graphical cues (Ericson *et al.*, 2004) or the likelihood of the target appearing from a direction (Kidd Jr. *et al.*, 2005) has been found to improve the performance in selective attention CRM tasks.

Research into the influence of the amount of angular separation between competing speech

sources has observed inconsistent effects on selective attention tasks with some papers reporting spatial unmasking independent of the amount of spatial separation (Nelson *et al.*, 1998), while others have found further benefits for greater separations (Ericson & McKinley, 1997; Spieth *et al.*, 1954) or less linear relationships (Shinn-Cunningham *et al.*, 2001). The angular separation is also dependent on the azimuthal position of the sources, with greater separation being required at the sides than when sources are near the median plane (Brungart & Simpson, 2003, 2005b).

In divided attention scenarios, improvements in CRM colour-number identification have been observed due to azimuthal separation with speech configured to remove any spectral overlap (Shinn-Cunningham & Ihlefeld, 2004). A later work by Best *et al.* (2006) also used speech modified to minimise energetic masking and varied the amount of spatial separation on the azimuthal planes. Best *et al.* (2006) found that colour-number identification in a CRM task improves with the amount of separation up to  $120^\circ$ , but only when ILD cues provide a better-ear advantage. A second experiment in the same paper in which the better-ear advantage was removed by the addition of noise, found that performance was improved for smaller spatial separations. Improved colour-number identification rates were also found when speech stimuli that had been modified to reduce spectral overlap, were spatialised and tested at various energy ratios created by varying one stimuli, with a constant call-sign, and leaving the other fixed at 70 dB (Ihlefeld & Shinn-Cunningham, 2008a). Further analysis of these results showed that the advantage was from improved identification of the quieter source, while the constant source was accurately identified independent of the spatial condition.

SRM has also been observed for concurrent speech with differences in elevation (McAnally *et al.*, 2002; Mesgarani *et al.*, 2003; Martin *et al.*, 2012). It is considered to be a smaller effect than for azimuthal separation (McAnally *et al.*, 2002; Mesgarani *et al.*, 2003). When combined with azimuthal separation it can have an additive effect (Mesgarani *et al.*, 2003), though not all studies have observed this (McAnally *et al.*, 2002). It has also been noted that SRM can depend on the relative elevations of the target and masking speech, as Martin *et al.* (2012) noted, changing the elevation of the masker did not effect intelligibility when the target was on the horizontal plane.

Brungart & Simpson (2001, 2002) also observed that separation in distance provided some SRM when concurrent talkers were of the same gender. This research led to the exploitation of distance cues in one of the designs proposed for the optimal seven-talker display in which five sources were positioned geometrically around the head [ $\pm 90^\circ$ ,  $\pm 30^\circ$ ,  $0^\circ$ ] at 1 metre and

two additional sources were positioned on the coronal plane at a distance of 12 cm (Brungart & Simpson, 2005b). When tested with seven competing talkers, out of the three possible designs, it was this configuration, referred to as *near-far*, that was found to perform best when better-ear normalisation was applied (Brungart & Simpson, 2003, 2005b).

As discussed in Section 3.2.2, when recreating spatial sounds over headphones it is possible to include all of the cues provided by the head through binaural processing which can be individualised if the listener's HRTF is used. Hawley *et al.* (1999) compared performance in a real sound field to that obtained using recording through a dummy head with concurrent speech and found no difference in intelligibility, though localisation was affected. A later work from Drullman & Bronkhorst (2000) found that there was no significant difference between the use of individualised or non-individualised HRTFs in terms of talker recognition, localisation and intelligibility in multi-talker scenarios.

Several papers have studied the difference between dichotic or intensity panning-based approaches and binaural presentations in multi-talker scenarios. Drullman & Bronkhorst (2000) showed that when three or four non-target talkers were presented to the ear that was contralateral to the target, intelligibility was comparable to when only one masker was presented in the contralateral ear and hence conditions with multiple talkers in the contralateral channel were omitted from the further analysis. The experiments, therefore, found that the binaural rendering resulted in significantly better intelligibility, talker recognition, and recognition response time than both dichotic and monaural presentations. The intelligibility interacted with the number of talkers and, for the one competing talker scenario, dichotic presentation performance was comparable with single-word intelligibility and higher with sentence intelligibility. It seems unlikely that the reduction in performance in the dichotic presentation is due to internalisation of the signals but, rather, is due to the lack of spatial separation between the target and the masking signals presented to the same ear. With this in mind, it is possible that spreading the additional talkers at intermediate points between the channels using intensity panning could facilitate performance that is comparable to those of the binaural presentation.

Studies which have investigated the differences between intensity panning and binaural presentations of multiple talkers appear generally to show an advantage of binaural presentation (Brungart & Simpson, 2005a; Carlander *et al.*, 2005; Kindström *et al.*, 2006). These studies, however, have either compared an optimised binaural presentation against a panning approach in which several signals could have occurred from the same location

(Brungart & Simpson, 2005a) or required participants to discriminate between two locations on the same side of the head (Kindström *et al.*, 2006).

### **Talker similarity**

One notable feature in naturally occurring multi-talker scenarios, is that all talkers' voices sound different due to the natural variation in the acoustic qualities of speech production systems. Timbral differences are difficult to classify, which makes the investigation of the exact features which make a 'good' combination of voices particularly difficult. This means although some studies report specific voices to be particularly well suited to multi-talker mixtures (Brungart, 2001), the qualities which facilitate this advantage are not readily apparent as is the case with more easily defined factors. A series of experiments have shown that mixtures of voices from different talkers cause less interference with each other than if the same voice is used for all sources (Brungart, 2001; Brungart *et al.*, 2001; Brungart & Simpson, 2007). It is thought that the unmasking observed when different voices are used is primarily due to a release from informational masking, as the different voices reduce confusion between which sounds belong to the masking or target speakers (Brungart *et al.*, 2009, 2001; Brungart, 2001).

Voices from different talkers are likely to vary both in  $F_0$  and timbre. Whilst the investigations mentioned up to this point generally simply used different talkers to create the timbral differences and consequently involved both spectral and  $F_0$  differences, Darwin *et al.* (2003) attempted to separate effects of the  $F_0$  from the alterations to the spectral envelope which they attributed to the vocal tract length. The spectral envelope was varied while keeping the  $F_0$  the constant through performing a pitch shift of the stimuli by factor  $vt$ , stretching the duration by  $1/vt$ , resampling the content at a sample rate of  $vtF_s$ , and then playing back at the original sample rate ( $F_s$ ) by ratios ( $vt$ ) of up to 1.34. A CRM experiment with two talkers was used to compare how the processing affected listeners' ability to attend to target streams. It was found that while small differences in  $vt$  made no difference to the probability of correct colour-number identifications,  $vt$  ratios of 1.13 or above significantly improved performance by as much as 17%. When the  $vt$  and  $F_0$  were varied together and found colour-number identification was improved more than by either  $F_0$  or  $vt$  alone.

It has been known for a long time that in dichotic presentations the sexes of the talkers are a significant factor (Treisman, 1964). As the number of talkers in a scene increases beyond two, it is no longer possible to assign talkers of different sexes to each stream. It has been



shown that in selective attention tasks when the target voice is of a different sex to the competing voices, performance is enhanced over that recorded with same sex mixtures or where competing talkers of the same sex are present (Brungart *et al.*, 2001; Joshi *et al.*, 2010; Shafiro & Gygi, 2007). Interestingly, it would seem that when one talker of a different sex is present, attention is drawn to that talker regardless of whether they are a target or competing source and therefore causes a lower performance than with same sex mixtures; a phenomenon referred to as ‘odd-sex distraction’ (Brungart *et al.*, 2001).

### **Onset asynchrony**

Differences between the timings of words spoken by different talkers (onset asynchronies/delays) may also help users to segregate and attend to different speech streams. Hedrick & Madix (2009) performed experiments using different vowel combinations with onset delays ranging from 0 to 150 ms in order to investigate the effect this had on double vowel identification tasks. All but two of the vowel pairs exhibited improvements in vowel identification performance with increasing onset delays. From the results, it was observed that the preceding vowel tends to suffer more from the masking and be most improved by the asynchrony, which was taken to suggest that the effect was due to informational masking. It is interesting to note the similarity of the findings with those associated with BRM (as seen in (Massaro, 1974)), which would be expected to exhibit similar trends to those observed here and accounts for the poorer performance of the leading vowel.

In a recent work, Fogerty *et al.* (2012) investigated the effect of ageing on the ability to determine whether one vowel was present, two were present but overlapped or two were present with a silent period between them. Three vowels were extracted from recording of the same male voice in /p/-vowel-/t/ contexts and resynthesised to have constant  $F_0$  (100 Hz). The stimuli were combined with different stimuli onset asynchrony (SOA) (0 - 175 ms) and presented monaurally. The results showed the gap-identification boundary was significantly lower for the younger group than for middle-aged or older groups, whilst the boundary for middle-aged and young groups was significantly lower than the older group for the judgement of the number of stimuli. Interestingly, the boundary between perceiving the mixture as one and two sources was very low (below 20 ms in all but one combination for the older group). It is expected that though the participants were able to determine the presence of more than one vowel, the ability to accurately recognise their identity would have only

become possible at much higher SOAs. Although this experiment does not deal with the SOA required for vowel recognition, the requirement for larger SOAs to determine the number of stimuli suggests that older users may also require larger onset delays when performing other speech based tasks (e.g., word recognition).

Assmann (1995) investigated the effect of formant transitions on the perception of concurrent vowels, initially with the theory that the formant transitions would contribute to the segregation of streams. In these experiments, a steady state vowel was combined with another syllable of the same length which was either: another steady state vowel, a consonant-vowel syllable, a vowel-consonant syllable or a consonant-vowel-consonant syllable. Concurrent vowel identification testing showed that the presence of a formant transition, either at the beginning or the end of one of the vowels, caused a significant increase in the number of correct identifications of the isolated vowel, whilst having no significant effect of the identifiability of the vowel with the transition. The lack of improvement in the identification rates of the vowel with the transition led to a further test in which the participants were given the identities of the two vowels and had to select the vowel in which the transition occurred or indicate the lack of a transition. Although the participants correctly identified the vowel with the transition in it more than incorrectly mistaking the steady-state vowel for it, when combined with inaccuracies in identifying the presence of a transition it was found that in most trials participants were unable to correctly recognise the vowel. This was taken as evidence against the significance of formant transitions in the grouping of features following Gestalt principles. Further research on this phenomenon has suggested that the observed identification improvements of the steady state vowel were simply due to the steady-state vowel's identity being less obscured during the transition (Assmann, 1996).

The effect of onset asynchrony above a sub-word level has however escaped the level of investigation paid to vowels, and consonant-vowel transitions. This is despite early observations with the effects of temporal overlap between competing speakers in radio communications (Webster & Thompson, 1954), in which it was noted that in 'divided attention' tasks with two speech signals presented, performance decreased systematically with the increase in overlap. However, analysis of the results found that the overall flow of information to the operator was greater than would have been achieved with serial display.

Best *et al.* (2011) explored the effects of varying the amount by which two short words overlapped (100, 50 and 0%) and whether they were presented from the different spatial locations or collocated. The experiment had two groups of participants, one that had

normal hearing and one in which the participants had sensorineural hearing loss. The results indicated that the group with normal hearing generally performed better in the 50% and 0 % overlapping conditions than the 100% overlapping condition when the sources were collocated. When the sources were separate, however, there appeared to be little effect from the amount of overlap.

Lee & Humes (2012) investigated the effect of onset asynchrony and  $F_0$  separation on performance for people of different ages and with normal and impaired hearing.  $F_0$  separations of 0, 3 and 6 semitones were introduced with preserved intonation from the original recorded speech, whilst onset asynchronies of 0, 50, 150, 300 and 600 ms were tested. Participants were split into four groups: young normal hearing, elderly normal hearing, elderly hearing impaired and young normal hearing with masking. The latter condition replicated the average spectral hearing response of the elderly hearing impaired group through the addition of shaped-noise to the stimuli. The experiment followed the selective attention procedure of the CRM test, with ‘baron’ as the target call-sign. In the first experiment, when an onset asynchrony was introduced the target would always precede the masker. Following stimuli presentation, the participants were asked if the target had been present and then to identify the key words via a touch-screen display. The results of the first question were referred to as the detectability, while the second question was referred to as identifiability and was scored correct only if both colour and number were that of the target. The results indicated that  $F_0$  separation and onset asynchrony facilitated improved target detection in all groups. With the identification results it was found that the groups performed equally poorly where onsets were synchronous and  $F_0$  were the same but exhibited different degrees of improvement as the differences were increased, with elderly hearing impaired group less able to take advantage of small onset asynchronies or pitch differences. It was found that a high proportion of the errors in all groups were due to intrusions—errors in which one or both words were from the masking stimuli—indicating that informational masking was a significant factor. It was found that the intrusion rate decreased as the differences in onset time and  $F_0$  were increased except in the case of the elderly hearing impaired group, where the number of intrusions increased between the 300 and 600 ms conditions. A follow-up experiment, in which the order of target and masker sentences were randomised for new participants with a reduced number of conditions (two  $F_0$  separations: 0, 3 semitones and two onset conditions: 0, 300 ms), found a significant but reduced advantage of the  $F_0$  or onset differences compared to the first experiment. The 300 ms onset asynchrony and 3 semitone conditions alone producing a similar result that was doubled when they were combined, which was taken as an indication of two separate systems

in action. As in the first experiment, the younger participants displayed significantly higher performance than either of the older groups.

### 3.3.4 Influences of visual information on speech perception

In Section 3.2.6, work was reviewed that demonstrates the links between visual and auditory perception. Speech perception is a particularly important case where links are known to exist. When listening to speech we are usually also provided with supporting visual information from the talker, such as their lip movements and gesticulations. The intelligibility of speech in noise is improved by being able to see the talker (Sumbly & Pollack, 1954; Middelweerd & Plomp, 1987; Helfer, 1997). Being able to visually attend to a talker is also known to have important influences on the perceived identity of speech. Seeing lip movements can alter a participant's identification of syllables (McGurk & MacDonald, 1976; Green & Kuhl, 1989; Green *et al.*, 1991), commonly referred to as the *McGurk effect*. This can be affected by the locus of visual attention. When presented visually with two talkers with different lip-movements, the perceived identity of a spoken consonant depends on the locus of visual spatial attention (Andersen *et al.*, 2009).

The “ventriloquism effect” (Howard & Templeton, 1966) (discussed in Section 3.2.6), where a visual stimulus alters the perceived location of a sound, is observed when speech is presented alongside a visual stimulus of someone talking (Witkin *et al.*, 1952; Thurlow & Jack, 1973; Bertelson *et al.*, 1997). There is some evidence that a visually perceived, spatially distinct talker can facilitate some spatial release from masking due to this, and there are benefits even when concurrent speech sources are collocated (Driver, 1996).

Visual attention has also been observed to impact the perception of speech. When simultaneous speech is present, visually attending to the location of the masking signal has been reported to have negative effects on participants' ability to recall the target stream (Reisberg *et al.*, 1981; Spence *et al.*, 2000). Furthermore, Driver & Spence (1994) observed that the benefit of being able to see the talker of attended speech is higher when the auditory and visual stimuli are co-located.

## 3.4 Summary

The transition from sound as an acoustic phenomenon to a perceptual experience is a complex process, not all of which is fully understood. This chapter provides a basic introduction to

important general principles of the process and factors which are known to be of particular significance to the topics of this thesis. While this thesis is not directly concerned with the physiology of the human auditory system (Section 3.2.1), it imposes important limitations on our perceptual capabilities. It is presented, therefore, to provide a basic understanding of relevant perceptual concepts.

Masking (Section 3.2.3) is a particularly pertinent example of these limitations. Both energetic and informational masking impose restrictions on the combinations of sounds that can be presented without interference. Within the design of auditory displays in which items overlap, or even occur very shortly before or after each other, these effects need careful consideration.

How sounds are organised at a perceptual level has implications for the design of auditory displays. Sounds being segregated incorrectly in an auditory display could result in information being lost and/or misinterpreted. Section 3.2.4 provides an overview of the relevant ideas of auditory stream analysis as put forward by Bregman (1990). This includes sequential and simultaneous, as well as both primitive and schema-based organisation. Additionally, attention (Section 3.2.5) plays an important role in our experience of auditory events. The ability to orient attention to specific frequency regions and locations in space allows for listeners to process specific elements of a busy auditory scene. These principles are of clear relevance to the design of all auditory displays, and particularly those in which users are confronted with simultaneous auditory sources.

The ability to perceive sounds as originating from locations in space is an important feature from an evolutionary perspective. This is clearly useful within the design of auditory displays, as the sections on the formation of auditory streams and factors affecting performance in concurrent speech scenarios both identify spatial locations as important cues. From understanding the cues that inform spatial location (i.e., ITD, ILD, interaural envelope difference (IED) and the response of the pinna), it is possible to simulate spatially positioned sounds over headphones. This opens up the opportunity for providing spatial auditory experiences without the need for complex, and expensive, loudspeaker configurations, which is clearly important when considering interfaces for consumer use.

Speech is a special type of sound for humans and, due to its ability to convey information, has clear relevance to the design of auditory displays. Sections 3.3.1 and 3.3.2 provides an introduction to speech signals and the process by which a listener recognises words. It seems likely that, regardless of whether speech is used, any auditory display will rely on a user's

ability to recognise and extract meaning from auditory codes and will, therefore, be based on common principles. Section 3.3.3 provides an in-depth review of the factors that affect performance in scenarios involving multiple concurrent talkers. From this review, it would seem that with any concurrent presentation of speech it is important to limit the number of concurrent streams of speech and present them from distinct spatial locations using voices with different fundamental frequencies, talker identities, and onset times.

This chapter has also highlighted the influences of vision, both generally on non-speech signals (discussed in Section 3.2.6), and on the perception of speech (discussed in Section 3.3.4). From this, it should be clear that auditory perception does not exist in isolation. Moreover, information from the visual system can have large influences on our understanding of a sound's location, and even its identity. This has crucial implications for the design of spatial auditory displays, as the location of a user's visual attention may influence the user's perception of the sounds intended to convey information.

## Chapter 4

# Auditory Display

In the first chapter, auditory displays are highlighted as a promising method for non-visual displays for television. The use of the auditory modality within HCI has a host of potential benefits. The addition of audio can (Peres *et al.*, 2008):

- enhance the communication of information to users who are visually-impaired or blind
- maintain the presentation of information to users whose visual attention is engaged elsewhere
- reduce the visual information that a user must process from the display at that time
- support the visually displayed information
- make an experience more affective
- present information from a device which has insufficient graphical output capabilities (i.e., no screen, low resolution display, or not enough screen space)
- draw a user's attention to errors

The field of auditory display is growing in maturity and a large variety of techniques for conveying information through sound have been proposed by various researchers. This chapter provides a detailed introduction to these techniques. Their individual merits and limitations and examples of displays that have been produced are discussed. The relevance of these designs to the use cases identified in Chapter 2, for menu-navigation and orchestrated synchronous companion experiences, are considered and outstanding research questions are identified which provide the motivation for the practical element of this project.

## 4.1 Serial speech displays

Since the primary aim of auditory displays is to communicate information through sound, it would seem obvious to use speech, as it has evolved specifically to allow information transfer and it is the most common means of human communication. In many ways, speech would appear to be the ideal solution for auditory displays. Speech utilises the language skills that the user has developed throughout their life and allows complex ideas to be conveyed, whilst at the same time allowing for a high degree of precision.

Serial speech interfaces display content through a single *stream* of speech. The most common example of these systems, and probably auditory interfaces in general, are *screen readers*. Screen readers are primarily used to enable access to computer systems for users who have visual impairments or who are blind. To accomplish this, these systems use text-to-speech (TTS) technology to convert textual content, such as on-screen text or metadata defined within mark-up descriptions, into speech. This allows the user to listen through the on-screen content and access most of the information available to users who are able to attend visually to the screen. The use of TTS technology in speech displays allows the display to be extremely flexible, capable of representing any information presented in a textual format. Steady improvements in the field of TTS synthesis mean that the quality of the speech in these systems is generally quite high, allowing information to be clearly communicated.

A disadvantage of screen readers, however, is that they are used to transform a graphical display into audio and therefore the resulting display is highly dependent on the manner in which the visual display is formatted. In GUIs, the spatial layout is often used to convey further information and relationships between objects, which the transfer to serial speech removes and can result in the meaning being lost (Raman, 1997). This poses a second issue in that the screen reader, and the systems that it represents are not a cohesive whole, which can cause confusion (Theofanos & Redish, 2003). Furthermore, for some source material it is practically impossible to create a full textual representation. Consider, for example, a painting. While it is entirely accurate to describe van Gogh's *Sunflowers* as: "a painting of fifteen sunflowers in a vase", it is clearly true that the description does not provide anything resembling the same amount of information that is present in the original piece. While this is a short description and further elaboration could be added, ("the background is yellow", "the vase is yellow and white"), capturing the full effect of the original is not possible. With this example, it would arguably never be possible to produce the same experience in another modality as actually seeing the original image. It may be possible, however, to provide a



more *equivalent* experience.

A major issue with accessing information in this manner is the amount of time that it takes to find information of interest. As pointed out by Kramer (1994, p. 47), speech “has a low information transmission rate for continuously changing variables relative to the bandwidth of the human auditory system”. Where the original visual interface presents the user with choices, they are forced to listen to each individual option to decide whether it is of interest or not and in more complex interfaces the amount of information can become large (Barnicle, 2000). This has a knock-on effect on the time taken to perform the navigation using speech. In order to avoid the large overheads in navigation time which are involved with speech displays, many users of screen readers configure their systems with very high speech rates (Theofanos & Redish, 2003; Borodin *et al.*, 2010). Some researchers have proposed to address the speed issues through serial speech auditory displays using accelerated speech rates, such as *SpeechSkimmer* which used different forms of temporal compression to allow users to skim through spoken content at different rates (Arons, 1997).

Up to this point, GUIs have been discussed as if they are static, but this is untrue in many situations. On many web services, pages update as new information becomes available or as the user interacts with them. Having only a single stream of speech means that users are unaware of updates, unless explicitly notified, and can be disorientated by changing content (Brown *et al.*, 2012). Furthermore, the cost of an irrelevant update may be considerable, as users are unable to make a quick glance to assess the change (Brown *et al.*, 2012).

Another issue is the load placed on the user’s memory, due to the inherent transience of speech (Pitt & Edwards, 1997). This issue is well demonstrated by another common speech interface: the automated telephone menu. The user is presented individually with each item within a sub-menu, requiring the user to remember the previous items in order to assess which option is most suitable. This can result in errors due to premature decisions being made before all options have been presented, or forgetting which of the responses mapped to the most suitable option after listening through the entire list. This is not such an issue in visual menus, as typically all items in a sub-menu are presented simultaneously and the user switches their attention between the options in order to decide which is most appropriate (Pitt & Edwards, 1997).

In recent years, speech-based displays have also permeated into many portable devices’ operating systems in the form of intelligent personal assistants (e.g., Apple’s Siri (Apple Inc., n.d.b) and Microsoft’s Cortana (Microsoft, 2016)). These systems are simple conversational

interfaces taking spoken natural language input and responding with visual and/or auditory information. It could be argued that the popularity of these services demonstrates the potential utility of spoken interfaces for the general population, rather than purely as an access tool for those with visual impairments or who are blind. While the conversational aspect may reduce the amount of visual interactions required, these systems face many of the same problems as screen readers. When there is a large amount of information to present or the information is not easily translatable into text a reliance on serial speech would, again, be problematic. Many of these systems, therefore, frequently defer to the visual display as output in situations where this is likely to be the case.

## 4.2 Non-speech auditory displays

Though most auditory interpersonal communication is spoken, in day-to-day life the auditory system uses information from a much wider array of non-speech sources to inform our understanding of our environment and the events occurring around us. It is often easy to overlook the wealth of information which can be extracted quickly from relatively simple non-speech sounds with minimal conscious effort. Consider a knock at the door, for example. At the most basic level, we are able to detect that there is a sonic event and identify that it is from someone knocking at the door. There is, however, more information encoded in this sound. The sound is linked to a meaning: ‘someone wants to get in’. Furthermore, it is possible to tell how hard they are hitting the door and from this infer information about the visitor and their motives. The rhythm of the knock may be familiar and reveal the identity of who is knocking. More complex arrangements of non-speech sound can be even more powerful. Instrumental music is a clear demonstration of how non-speech sounds can be organised to communicate information. With instrumental music, this information is often emotive and abstract. It is, however, also able to be associated with other concepts to convey or support a narrative. Notable examples of this include the use of themes in Prokofiev’s (1936) ‘Peter and the Wolf’, and Herrmann’s soundtrack for ‘Psycho’ (Hitchcock, 1960).

It is clear that limiting auditory display methodologies to speech-based representations fails to utilise the full capability of the human auditory system. An obvious analogy is the use of non-textual information in visual interfaces, in which similar issues are experienced (Blattner *et al.*, 1989). In fact, this use of non-textual information has been such a success in improving the usability of visual interfaces that most operating systems’ user interfaces are full of examples. Given the capabilities of non-speech audio, designing purely speech-based auditory

displays is akin to designing visual interfaces as plain text.

The use of non-speech audio has seen a considerable amount of development over recent decades. From the research community, several distinct methodologies have emerged which are outlined and discussed in the remainder of this chapter within the context of television displays.

### 4.2.1 Audification, parameter mapping and model-based sonification

When trying to gain an understanding of complex data, it is common practice for a scientist to visualise the data to gain an understanding of the data's structure. In this form, it may be possible to notice patterns or trends that would be hard to identify from looking at the raw values alone. Audification and sonification offer similar advantages but in the auditory domain and exploiting the source analysis and pattern recognition elements of auditory perception.

Audification refers to interpreting data “as amplitude over time and playing it back on a loudspeaker for the purposes of listening” (Dombois & Eckel, 2011, p. 301). While some datasets may be audible in their raw state, audification often involves signal conditioning in which preprocessing, such as time and/or frequency manipulations, are applied to the data signal so that any features of interest may be detected by the auditory system (Dombois & Eckel, 2011). This approach, however, has implicit limitations on the type of data that it can represent. The data must be transformable into a time-varying signal that may be interpreted as a sound wave and consist of a large number of samples so that it may be analysed at audio-sampling rates (Dombois & Eckel, 2011).

Sonification is further split into two subcategories: parameter mapping sonification and model-based sonification. Parameter mapping sonification is a methodology in which the data is used to modify parameters of sounds produced by the display (Grond & Berger, 2011). This approach typically involves the data being used to modify parameters of a synthesiser (e.g., oscillator frequency, waveform mixture, amplitude) and/or an effects unit. Within this model, the information does not create the sound, but manipulates its characteristics. In model-based sonification, the data is used to create a model of a system that can be interacted with to produce sounds (Hermann & Ritter, 1999). The approach effectively involves the construction of a virtual instrument from the data itself, which the user can excite as they choose to explore the data (Hermann, 2011).

Both audification and sonification are particularly well suited to investigating aspects of

complex datasets. This has led to suggestions of their use for analysing seismic recordings (Hayward, 1994), electroencephalogram data (Hermann *et al.*, 2002) and financial data (Janata & Childs, 2004). Though it is true that all information can be represented numerically and, therefore, represented using these techniques (Hermann, 2008), it is likely that the resulting sonification would be exceptionally complex and require a considerable amount of training to reliably interpret. In scenarios where auditory representation of information is required, as is the case in many television displays, these approaches are unlikely to provide a useful alternative to spoken representations.

### 4.2.2 Auditory icons

Gaver (1986) proposed the first design methodology that exploited the semantic potential of non-speech sound, which he dubbed *auditory icons*. Gaver noted the way in which people use sounds to gather information in the real world, separating the *proximal* cues (regarding pitch, duration etc.) from the *source* (the material, type of excitation, etc.). He highlighted the ability to determine characteristics of a source from *everyday sounds* and proposed the manipulation of natural sounds, in terms of the characteristics of the source, would better utilise human auditory abilities than modifications based on musical features. Gaver suggested that in an auditory display for a computer, through manipulations of source properties, it would be possible to convey added information regarding the item being represented, such as its size or location. In addition to using sounds that were “caricatures of natural sounds” (Gaver, 1986, p. 173), Gaver also recommended that, where possible, sounds be chosen that had either a direct causal link (*nomic*) or a *metaphorical* link to the item being represented (e.g., representing a camera with the sound made by a shutter release). The idea of these requirements was that a clear semantic link between sound and object or action would ease learning for users.

In a later work, Gaver (1989) described an implementation of auditory icons in the *Finder* interface on Apple Macintosh computers in a system referred to as *SonicFinder*. This system was used alongside the graphical display by several users over the course of a year. Although no formal user testing was performed, it was noted that users reported missing the sounds when they were not present and Gaver referred to the experience as being more engaging. It was speculated that “increases in speed or accuracy associated with the addition of sound to this interface seem likely to be small or none” (Gaver, 1989, p. 84).

Auditory icons can take many different forms depending on the decisions of the sound

designer. The auditory icons in *SonicFinder* supported the perception of the user that they were interacting with the tangible items suggested through the terminology and visual design of the pre-existing interface (Brazil & Fernström, 2011). Other authors have created interfaces which have significantly departed from traditional metaphors to create more naturalistic auditory environments (Mynatt *et al.*, 1998).

In (Mynatt, 1994a), factors affecting how effective auditory icons would be are identified as: how *identifiable* the sound was, the strength of the *conceptual mapping* between sound and content, the properties of the audio clip used (length, sample rate, bandwidth), and users' individual emotional responses to specific sounds. Arising from these observations, a number of guidelines were suggested for the design of auditory icons and tested (described in (Mynatt, 1994b, a)). These guidelines stipulate that auditory icon designers should choose short, wideband sounds with consideration of their discriminability. They should then run through a series of experiments to ensure the stimuli are sufficiently identifiable or, failing that, easily learnt (Mynatt, 1994a). Finally, they should evaluate possible mappings to the items and perform usability testing (Mynatt, 1994a).

In a further variation of the auditory icon, referred to as the 'parametric auditory icon', aspects of the icon are manipulated to convey further meaning (Brazil & Fernström, 2011). Stevens *et al.* (2004) investigated the potential of parametric auditory icons by manipulating cues such as the spatial location, pitch, reverberation and filtering to indicate properties such as size, direction and distance. They found that recognition can suffer as the number of parameters is increased.

Regardless of whether parametric or standard auditory icons are used, it seems as though the design of any interface based on auditory icons requires a considerable level of involvement from a sound designer and a significant amount of user testing for each additional cue added to the interface. Both of these are costly implications for any company considering the development of an auditory interface. These restrictions also severely limit the feasibility of the deployment of auditory icons within scenarios in which the representation of dynamic content is required (Walker *et al.*, 2006, 2013).

### 4.2.3 Earcons

The concept of *earcons* was proposed by Blattner *et al.* (1989) as an alternative non-speech cue to the auditory icon. Gaver (1986) outlined three potential mappings for non-speech cues: nomic, metaphorical and symbolic, and proposed that auditory icons should be chosen from

natural sounds to provide iconic or metaphorical mappings. By contrast, Blattner *et al.*'s (1989) earcons were musically-based and therefore inherently abstract with no pre-existing semantic link between the sounds and the content that they represented. According to Gaver's (1986) semiotic categories, this makes them a symbolic representation, as they rely entirely on learnt associations with their associated content. Although this abstraction reduces how intuitive the auditory-item pairings may be, the use of musical cues allows a greater degree of complexity and, consequently, a much larger amount of information can be conveyed. In fact, although no semantic link is considered during the design of the earcons, user feedback reported after recognition experiments suggests that people form their own semantic links between the cues and the represented object, which therefore implies a greater similarity between the auditory icon and earcon cues than theoretical consideration would suggest (Brewster *et al.*, 1998).

Although using abstract auditory representations may seem likely to confuse users, it is useful to consider a common system that relies on learnt associations of sounds with otherwise unrelated concepts; the phone. McGookin & Brewster (2011) referred to the use of different, user-selected ringtones for different callers as an example of users learning arbitrary associations between sounds and information, in this case the caller's identity. With the multi-functionality of modern phones, it is now common that a larger number of different audio notifications are used. This use of sound means that, from the notification sound alone, the user is aware of whether they have received a text, email, social-media update or a phone call. The mobile phone, therefore, continues to be a clear example of the potential utility of earcon design concepts.

In addition to the basic use of individual earcons, *compound* earcons were also suggested, where two or more earcons are concatenated to form longer phrases with more information (Blattner *et al.*, 1989). For example, in a television interface one may combine earcons for 'film' and 'comedy' to denote a comedy film. Another interesting type of earcon suggested by Blattner *et al.* is the *inherited* earcon, in which common musical features (timbre, rhythm, pitch, register) are used to denote the family and identity of a specific earcon. Through having a common motif (rhythm) and altering the timbre, dynamics, register and pitch, the interface may convey broad information about the category of item being represented and fine detail to identify it further (Blattner *et al.*, 1989). Inherited earcons have been commonly applied in hierarchical menu structures to provide location information (Brewster *et al.*, 1996; Vargas & Anderson, 2003). They can be combined with spoken menu representation by prepending

them to the TTS synthesised labels (Vargas & Anderson, 2003).

For earcons to be an effective cue, they need to be memorable and distinctive to ensure that users are able to easily tell them apart and, over time, to learn them and their associations. Being based on musical concepts, it is plausible that the ability to learn and recognise earcons is dependent on musical experience and ability. Experimentation with musicians and non-musicians, however, found that although musically-trained participants performed better when simple signals such as sine, triangle and square waves were used, no significant difference in earcon recognition was found when instrumental timbre differences are also included (Brewster *et al.*, 1993). In a later study, some non-musicians were still found to have exceptionally low recognition rates ( $< 65\%$ ) after being trained with the earcon set, which was attributed to tone deafness (Brewster *et al.*, 1995b).

Experimentation into the use of earcons has resulted in a set of design guidelines (Brewster *et al.*, 1995a), which specify:

- the use of instrumental timbres
- avoiding the use of register as an absolute cue unless large differences (2-3 octaves) are used
- avoiding excessively high or low pitches
- maximising the rhythmical differences between earcons
- keeping intensities of an earcon set within a small range
- using spatial separation
- making the earcons attention-grabbing and leaving short gaps between serial earcons.

Earcons offer a distinct advantage in that, as they are not bound to capturing (or synthesising representations of) natural sounds, it is possible for them to be synthesised within the system. This offers a reduction in the amount of storage required, which is important for smaller devices, but also introduces the possibility of an automated earcon design where a set of inherited and experimentally derived perceptual laws would constrain a generative music program. Even if automatic generation were possible for hierarchical earcons, however, a small change in the structure of the represented system would require the user to re-learn a large quantity of the earcons (Walker *et al.*, 2006, 2013).

A further development to the earcon concept was proposed called morphological earcons or *morphocons* (Parseihian & Katz, 2012). These cues remove the earcon's restrictions on timbre, so that they instead solely rely on the temporal manipulation parameters to construct the motifs. This move away from timbral cues was primarily to allow users to change sound-palettes, while minimising the amount of training required to re-learn the stimuli. Rather than using MIDI (musical instrument digital interface) files to produce musical modifications as with earcons, samples were modified using time-stretching and pitch-shifting algorithms. Identification testing with sighted, visually-impaired and blind participants found that the icons performed well overall. As no comparison was made with performance using traditional earcons, it is not possible to analyse the effect the timbral independence had on the stimuli's identifiability. From observations in user testing (Brewster *et al.*, 1993) and personal experience of listening to earcons, however, it often seems that timbral differences act as a key difference between the stimuli, which becomes apparent much sooner than the temporally dependent features manipulated for the morphocon cues. It therefore seems that, although the timbral independence may allow a reduced learning time when changing between sound palettes, by avoiding the use of timbre to differentiate objects the flexibility of the cue would be limited.

Interestingly, Parseihian & Katz (2012) do not suggest the use of iconic sounds within the palette to create a fusion of the auditory icon and earcon cues. This suggests a reliance on timbre and therefore is a considerable shift from the motivation of their paper. As the modifications used in the morphocon creation are applicable to most audio signals, there is no apparent reason why the cues could not be applied to *natural* sounds. This could facilitate the improved learnability associated with semantically-linked sounds and the ability to convey similarities between items or location within a hierarchy.

#### 4.2.4 Musicons

McGee-Lennon *et al.* (2011) proposed the concept of the *musicon*. It exploits our ability to recognise snippets of familiar music and constructs an alternative language through the use of short clips of familiar tracks. Musicons, like earcons, rely on learnt bindings between the cue and the represented item. McGee-Lennon *et al.* (2011) argued, however, that familiar music would have stronger responses and consequentially be more memorable. These cues were designed initially to be used as reminders and a key part of the justification focussed on the need for notifications whose meanings were clear only to the intended listener, which



the abstract representation facilitated.

A variety of musicon lengths were tested both directly after training and a week later with no additional training (McGee-Lennon *et al.*, 2011). Though good performance was observed with sounds as short as 200 ms, the best compromise between response time, user preference and accuracy was found for a duration of 500 ms. The accuracy observed with the musicons was high, but lower than for spoken reminders which, unsurprisingly, achieved 100% recognition. This approach, however, introduces an issue from the perspective of a company wishing to deploy such a display. Due to rights issues, it is unlikely that circulating a display with recognisable and, therefore, commercially successful music included would be feasible. It would, therefore, be necessary for a display to rely on a user's existing digital music collection to create stimuli. Further work on musicons by McLachlan *et al.* (2012) has indicated that it is feasible to allow participants to choose their own tracks and sections from which to create musicons. Although a small reduction in recognition rate was observed, they were preferred by the participants (McLachlan *et al.*, 2012). In practice, this level of customisation means that it is likely that no two users would have the same experience, which would be problematic in scenarios in which many users may need to use the system.

#### 4.2.5 Spearcons

Walker *et al.* (2006) proposed a new type of non-speech cue that differs considerably from those reviewed so far, which they called the spearcon. The innovative concept behind the spearcon is the exploitation of the acoustic complexity of spoken words, which have developed in such a way as to be distinguishable from one another. Spearcons leverage users' pre-existing language skills to semantically link cues to content, whilst minimising presentation times. To achieve this, speech is synthesised from the content labels using a TTS synthesiser. Time-scale modification is then performed using a pitch-constant overlap-add algorithm to reduce the duration of the speech by as much as 50%. The reduction of the duration of the speech is to allow faster navigation whilst maintaining the *recognisability* of the speech rather than its *intelligibility*. Although a spearcon's meaning may not be apparent initially, after a short amount of training it should be easily recognised. This, therefore, defines the cue as being non-speech despite originating from speech materials and distinguishes it from the sped-up speech approaches mentioned in Section 4.1.

As the spearcons are produced via TTS, it is feasible that they could be implemented within a system dealing with dynamic content without the requirement for a sound designer.

Furthermore, due to the strength of the semantic link and the distinctive nature of individual spearcons, the spearcons require significantly less learning effort than auditory icons or earcons (Dingler *et al.*, 2008; Palladino & Walker, 2007; Walker *et al.*, 2013).

A question that is raised by the use of spearcons is the degree to which users rely on the intelligibility of the cues. Studies have previously shown high intelligibility rates of time-compressed speech for up to 50% compression rates (Fairbanks *et al.*, 1957a). Intelligibility drops rapidly as compression rates are increased further (Fairbanks *et al.*, 1957b; Heiman *et al.*, 1986). This implies that some spearcons are intelligible. There is some support for this from Dingler *et al.* (2008) who, in addition to comparing the learnability of the spearcons to non-speech cues, also compared the results to those achieved with unprocessed speech and found no significant difference between the two cues after one round of initial training. This appears to call into question the degree to which participants rely on the intelligibility of the signal. A follow up experiment, however, compared spearcon recognition performance of naive and trained users (who had completed an experiment using spearcons) and this indicated that although naive participant performance was high (averaging at approximately 64%), the performance of the participants was still significantly enhanced by the training (Palladino & Walker, 2007; Walker *et al.*, 2013). Further to this, in a study in which spearcons were created in German, initial recognition rates as low as 12% were reported (Wersényi, 2009).

Spearcons were also compared to earcons, auditory icons and speech in regards to their effects on navigational speed and accuracy in auditory menus (Walker *et al.*, 2006, 2013). It was found that the spearcon cue was faster and more accurate than other non-speech cues (Walker *et al.*, 2006, 2013). Research into how navigational speed in spearcon-enhanced menus compares to normal TTS menus found no improvement in one-dimensional menus (Palladino, 2007; Palladino & Walker, 2008a). In two-dimensional menu structures, however, spearcon enhancement improvements have been found (Palladino & Walker, 2008b), although not all experiments have shown this improvement to be significant (Walker *et al.*, 2006). A potential factor to explain this difference is slightly different amounts of time-compression used by the two approaches, whilst Walker *et al.* (2006) applied between 40 and 50% reduction in duration, (Palladino & Walker, 2008b) used a logarithmic function based on the length of the input speech so that long phrases were compressed more than short phrases.

As a relatively new type of cue, there are elements of the spearcon which have still to be investigated and are not fully understood. Firstly, as already mentioned, is the degree to

which the user relies on the underlying intelligibility of the signal and hence the extent to which spearcons are processed as speech by the brain. Until this is understood, it is hard to know whether the spearcon should be truly considered as a speech or non-speech cue. Also, the influence of factors such as the temporal compression algorithm and vocal parameters of source speech (gender, intonation, accent, prosody, etc.) are still not understood (Yalla & Walker, 2007). Potentially, these could be used to encode further attributes in the cue or to reduce masking effects when presented with other sounds.

#### 4.2.6 Auditory scrollbars and Spindex

When navigating large documents or lists on computers, it is easy to overlook the information we gain from the presence of the scrollbar. Through its position and size, it is possible for the user to estimate their relative position and how much content is present. Yalla & Walker (2007, 2008) recognised the importance of this information and began research into methods for the creation of *Auditory Scrollbars*. The auditory scrollbar, like the previously discussed cues, is prepended before each item description. It relates only to the position of the item within the list, which is represented by the pitch of the tone. By prepending the cue in this fashion, the user can move at considerable speed, hearing only the auditory scrollbar cues which forms a glissando representing the movement of the user through the list.

By experimenting with several designs, Yalla & Walker (2008) found that a *proportional grouping*, where the list is split into eight sections represented by the eight tones of an octave, was best in terms of preference and in the ability of users to estimate the size of the list. Proportional grouping means that the number of items represented by the same pitch indicates the total number of items in the list. The direction of the glissando for moving through the list appears to be a matter of preference (Yalla & Walker, 2008).

An alternative system was proposed under the name *Spindex*, which used a prepended speech cue to assist fast navigation in alphabetic lists (Jeon & Walker, 2009, 2011). The cue was simply formed of the first letter of the item name followed by a short gap before the item name. This allowed the user to understand their location within the list relative to the item that they were looking for, enabling them to move quickly to the appropriate general region and then to move more carefully to locate the target. This was found to reduce the amount of time taken by users to locate target items in larger lists (Jeon & Walker, 2009, 2011).

Repeating the first letter of each item could be quite repetitive when navigating through larger lists. A later experiment compared several different schemes in which only the first

item starting with a new letter was signalled and the spindex cues were either silenced or attenuated for subsequent items starting with the same letter. Results showed that users preferred the designs in which the repeated cues were attenuated either to a constant level or progressively (Jeon & Walker, 2011).

Both of these cues presume that the user has a good idea of the location of the target cue within the menu. The auditory scrollbar, as with visual scrollbars, provides an approximate representation only of the amount of information in the display. It chiefly benefits experienced users who know where the information or item of interest is located in the display. The spindex approach also requires alphabetic ordering, which is not appropriate for many types of interface.

#### 4.2.7 Discussion

The development of non-speech methodologies has opened up new means for presenting a user with information through audio which has less temporal redundancy than if speech were used. All of the cues require some degree of learning for the user to understand the links between particular items and their sonic representations. This imposes a limit on the number of effective non-speech cues; in part determined by the flexibility of the non-speech cue in order to allow cues to be distinguishable, and by the ability of the listener to remember the cues. User studies into the number of non-speech cues which can be reliably memorised would, however, require long studies and be highly dependent on the sounds used and the precise characteristics of how they were designed. These factors are not problematic when the number of items to be represented is small and little content regularly changes, but it is likely that many television interfaces will have a large quantity of frequently changing content. As spearcons are constructed from spoken cues that can be dynamically generated using TTS systems, it is likely that they would suffer less from these issues than other cues, both due to their intrinsic linguistic distinguishability and the user slowly learning to understanding the speech at the accelerated rates.

The non-speech methods described, with the possible exception of the spearcon, can be viewed as attempts to create or use existing non-speech vocabularies to form new languages. In fact, it may be argued that the advantage of non-speech cues is due to the limited number of possible meanings and therefore the reduced complexity required to distinguish them from other cues. Whilst speech communication is limited in how it can represent information by the physiology of the speech production system and, hence, uses temporal ordering to achieve

the required flexibility, non-speech methods can draw from an infinite gamut of sounds. This allows a wider range of ideas to be represented with little need for structural cues, which require longer presentation times.

Recent investigations into the use of spearcons appear to indicate that these cues have considerable advantages over other non-speech cues (e.g., Walker *et al.*, 2013). In a working system, however, it is likely that many types of non-speech cue would be used to help distinguish object types and to best represent different categories of information, as proposed in (Wersényi, 2009). Also, it would seem that despite the advantages of the various non-speech methodologies, their inherent inability to convey novel information means that any audio interface is likely still to require at least some spoken information.

### 4.3 Concurrent auditory displays

The discussion so far has considered only the sequential presentation of auditory elements. Sequentially presenting information leads to a “presentation rate bottleneck” (Walker *et al.*, 2001, p. 532) in the flow of information from the system to the user, which imposes limitations on the speed with which a user may access specific elements of the information. The approaches outlined in the previous section explored how to extend auditory displays beyond the serial-speech interface by presenting shorter cues with the meaning encoded through different design methodologies. These approaches effectively seek to increase the efficiency of the messages passed through the bottleneck by removing redundancy associated with the spoken representation. A serial presentation, however, imposes a strict order onto the information that is presented. This removes the user from the decision about what information is most important to them at any given time. Due to the inherently temporal nature of the medium, some ordering will always be present in information presented as audio. The impact of ordering on the user experience, however, may also be reduced by presenting information concurrently.

The approach of presenting large amounts of information to the user and then allowing them to switch their attention between different items potentially has some interesting advantages. Firstly, switching attention requires no explicit interaction with the interface, therefore reducing the amount of interaction required and the cumulative effect of individual response times. Also, by increasing the number of items displayed at once, important content is not delayed due to the time taken to present other, less-relevant items (McGookin & Brewster,

2006). This potentially means that a concurrent display would be faster and more efficient to use. Furthermore, as the user is presented with more items concurrently, it is easier for them to remind themselves about useful information such as the other available options, which reduces the requirement to remember items compared with a serial display (Parente, 2008). Alternatively, it can inform the user of the context of the display such as on-going tasks (Gaver *et al.*, 1991), or their location within a system structure (Lorho *et al.*, 2002). Therefore, concurrency appears to considerably enhance the capabilities of auditory displays. However, for an auditory display to be successful it is necessary that the sources are perceived as separate and recognisable, with their informational content intact. Concurrency introduces additional difficulties, however, due to the limitations of human auditory perception. The different auditory representations, introduced in the last section, are likely to face different sets of challenges in concurrent presentations due to their different acoustic characteristics and the cognitive processing required to interpret them. Careful consideration must be given to the design of these displays. A number of researchers have already investigated the potential of concurrent presentations for a variety of cues and these are introduced and discussed in the following sections.

### 4.3.1 Concurrent earcons

In some ways, earcons and any auditory display could be perceived as music, as they require the organisation of sounds or at least the design of systems to organise sounds. For the sake of this work, earcons are differentiated from music in that they are entirely bound to their function of communication. However, as earcons are closely related to instrumental music, it would seem that they might also be suitable for concurrent presentation, provided they follow similar compositional restrictions. In Blattner *et al.*'s (1989) original work on the subject, the earcon cue was proposed primarily as a method for serially presenting information, but a suggestion was also made about using multiple earcons concurrently. These initial investigations ran into difficulties with the interactions between concurrent cues, which was again found in (McGookin & Brewster, 2002). Concurrent earcons replicate ideas from polyphony in musical theory, where multiple melodies are presented concurrently.

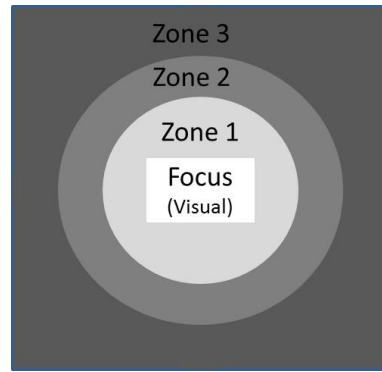
The difficulties described by McGookin & Brewster (2002) led to work identifying methods for the successful presentation of concurrent earcons using experimental techniques (McGookin & Brewster, 2003, 2004b; McGookin, 2004). The result of this work led to an additional set of guidelines for the creation of concurrent earcons (McGookin & Brewster, 2004b, a). These

stipulated that earcons should be designed such that each concurrently presented earcon had a different timbre, earcons had inharmonic relationships and considerable onset asynchronies, and spatial separation (McGookin & Brewster, 2004b, a). Even when adhering to these guidelines, however, the number of concurrent sources was restricted, with large reductions in identification performance as the number of earcons was increased to four (McGookin & Brewster, 2004b).

The limited number of concurrent sources appears to be at odds with the comparatively large number of instrumental lines common in musical compositions. However, whilst in music it is often intended that multiple instrumental lines are heard as one, in order for earcons to be reliably recognised they must be perceived individually. Also, in music, information is usually conveyed by a combination of all concurrent items, but for earcons to fulfil their purpose they must succeed in conveying their individual messages. It is, therefore, unacceptable for interactions between concurrent earcons to negatively affect a user's ability to distinguish the defining features of the constituent earcons.

As part of work into the use of design patterns to develop auditory interfaces, an interface for Microsoft Explorer was proposed which presented spoken objects and earcons within a virtual room (Putz, 2004; Frauenberger *et al.*, 2005b, a). On one wall was displayed the menu system and on another a *toolbar*. The menu was displayed using speech combined with instrumental tones, which were configured so as to increase in pitch from the first to last items—similar to the auditory scrollbar (Yalla & Walker, 2008)—and it was displayed across the virtual wall so that the structure originated in the bottom left corner and spread to the top right corner. The tool-bar was represented using five earcons positioned in a row along the other wall and additional background sounds were used to represent changes in state (such as the opening of a pop-up windows). Spatialisation was performed using binaural rendering and a head-tracker to improve localisation.

While the spoken menu items were triggered individually, the tool-bar earcons were presented concurrently to the user (Frauenberger, 2013). The display of the tool-bar system was configured such that initially the display was silent, but as the user moved towards a wall using a joystick, five concurrent earcons were presented. This is notably above the four earcons found by McGookin & Brewster (2004b) to significantly reduce users' performance. The interface was evaluated by Frauenberger *et al.* (2005b) using two groups of participants, either sighted or with differing degrees of visual impairment. After a short training session users responded favourably toward the system and made very few mistakes in selecting



**Figure 4.1:** Representation of McGookin & Brewster’s multi-modal display showing the focus and priority zones. The shade of the zones represents the increased importance required for a source to be sonified further from the focus. (Adapted from McGookin & Brewster (2001, p. 3))

folders or files. It was found that visually-impaired users favoured the tool-bar, while sighted users used the menu and tool-bar elements more equally. It is unclear what would cause this difference between the groups. The preference indicated by participants with visual impairments, however, suggests that this configuration of earcons was relatively easy to use. Unfortunately, although this work suggests that presentations of five earcons are possible, it is difficult to isolate the requirements for this.

McGookin and Brewster proposed an interesting solution to the navigation of large amounts of information distributed in two dimensions with their *Fishears* (McGookin & Brewster, 2001) or *Dolphin* (McGookin & Brewster, 2002; McGookin, 2004) systems. The systems were designed to represent a map of a theme park with a large number of rides, each represented by an earcon. The system used binaural rendering to spatialise the earcons in appropriate locations. To reduce the perceptual and computational loads, not all the rides were presented concurrently. Instead, the user could shift a *focus* zone that could be moved across the map, which was represented visually on the screen of a personal display assistant (PDA). A set of concentric *priority* zones were defined around the *focus* zone, which would sonify any rides within them which were deemed above a given importance level, with zones closer to the central focus having lower thresholds (see Figure 4.1). The experiment compared route-finding performance between a scrolling visual map and this multi-modal display. No differences were uncovered between the displays in terms of performance, although subjective evaluations ascribed a higher mental workload and higher levels of frustration and annoyance to the multi-modal interface (McGookin, 2004). McGookin (2004) noted that the earcons used in this experiment are unlikely to have been optimal due to their acoustic similarities, which resulted from the common ‘grammar’ used in their construction. Furthermore, it is



notable that there was no exploration of this design as a model for an entirely non-visual display, or a comparison with alternative auditory display methods.

### 4.3.2 Concurrent auditory icons

Brazil & Fernström (2006) and Brazil *et al.* (2009) proposed the use of concurrent auditory icons and tested identification rates with an onset asynchrony of 300 ms. In these experiments, upwards of three concurrent ‘everyday sounds’ were presented diotically to users, who were asked to identify as many of them as possible using free-text responses. Results indicated that much higher correct identification rates were achieved with concurrent auditory icons than had been the case for concurrent earcons (Brazil & Fernström, 2006). It was found that performance was improved when the sounds were selected on the basis of not being produced by the same *object* or *action* (or method of excitation) (Brazil & Fernström, 2006; Brazil *et al.*, 2009). This is logical from both informational and energetic masking perspectives, as the change in excitation method and material would facilitate significant spectro-temporal differences and so reduce spectral overlaps and timbral similarities.

Though the identification rates reported by Brazil & Fernström (2006) and Brazil *et al.* (2009) were high, it is questionable whether a comparison with the results using concurrent earcons (i.e., (McGookin & Brewster, 2004a)) is justifiable, as the two studies required participants to perform slightly different tasks. With the auditory icon experiments, the user simply had to identify the nature of each of the sound sources. In the earcon experiments, however, it was necessary for participants to recognise the nature of the sources and recall the associated content. This difference introduced additional complexity in the latter experiment, which would be expected to negatively affect performance.

Several displays have made use of concurrent auditory icons in different ways. Gaver *et al.* (1991) created a display to be used with a simulated factory. Up to 14 auditory icons were presented concurrently alongside a visual display, which could only display a limited section of the factory. The auditory icons were designed so as to have distinctive spectro-temporal characteristics to reduce masking and maximise discriminability. The display represented the machines using repeating loops of everyday sounds and used additional sounds to represent issues occurring within the factory (i.e., breaking glass and liquid spilling). Although recognition was not quantitatively measured, observations of participants working collaboratively in pairs on the simulation showed that they communicated more to solve problems when the auditory cues were presented, and responded reliably to error noises, but

often failed to notice when a looped sound stopped. Similarly, in one version of the *Audio Aura* display, proposed by Mynatt *et al.* (1998) to act as a peripheral display for the office environment, auditory icons were combined to create a soundscape of a seaside environment with different sonic elements representing different ideas.

Whilst both Gaver and Mynatt's systems served primarily as displays indicating states, Putz (2004) and Frauenberger *et al.* (2004) proposed an interface which exploited spatial separation with concurrently presented auditory icons to represent hierarchical menus. The auditory icons were presented in the frontal hemisphere on the azimuthal plane in a virtual room. The users were able to *zoom* in or out of the display which controlled the number of concurrently presented items using a Gaussian window. At the most zoomed out, the user was able to hear all items (although more attenuated at the sides) and at the most zoomed in only one item was audible (Putz, 2004). The user navigated through the menus by turning their head towards the target item to select it and then used a keyboard to perform the required actions. The auditory icons were looped but, in order to allow better localisation, different frequency tones (referred to as *pedestal tones*) were added to each of the auditory icons using a small amount of amplitude modulation. Putz (2004) reported that many of the participants kept the display set to maximum zoom or used a hint button repeatedly until they found the item that they were looking for. These findings indicate that users struggled with the concurrent presentation of the auditory icons. As the users had the ability to change the zoom, it is unclear whether they would have become accustomed to this mode if fewer items had been presented concurrently or if they had got more used to the concurrent presentation.

It would seem, therefore, that there have been some contradictory findings in terms of the number of concurrent auditory icons that can be used reliably. It is unlikely, however, that many more than three would be sensible in most task-orientated scenarios (e.g., menu navigation). A much larger number of concurrent stimuli is possible in ambient state monitoring displays, where the use case involves monitoring the states of continuous processes, as in Gaver *et al.*'s (1991) ARKola simulation. This is probably due to different listening techniques being required by the two scenarios. Firstly, in ambient state monitoring displays, the sounds are present for an extended period, in which time the user is able to switch attention between concurrent streams and gain familiarity with the stimuli. In task driven presentations, however, it is likely that the stimuli will be considerably shorter and only briefly displayed to the user, giving them little time to familiarise themselves with available options or switch attention. In the task-orientated scenarios, the user must focus on each

item individually, make a decision about the nature of the source and infer the object with which it has been associated. The state-monitoring display requires the user to listen to the timbre of the mixture rather than each individual sound. Then, when a change occurs due to the addition or removal of a source, the user has to determine the nature of the item which has changed.

### 4.3.3 Concurrent speech displays

As previously outlined in Section 4.2, speech contains temporal redundancy which makes it slow for navigating large amounts of content. Although these arguments have led to the development of non-speech methods, there are scenarios in which the occurrence of unfamiliar information, the number of possible options, or the familiarity of the user with the interface make a purely symbolic representation unsuitable. An interesting interface design involves displaying several concurrent speech signals from which the user is able to selectively attend to a desired piece of content. This model allows the desired content, whatever it may be, to be displayed to the user with minimal delay and without requiring the learning of non-speech cues. In a serial presentation paradigm, the target speech is presented in isolation with minimal competing noise. By contrast, in the parallel display, the target phrase has to compete with other speech signals.

Much of the work on concurrent speech has been concerned with the design of communications systems for time-critical, multi-talker scenarios. This work has contributed greatly to understanding the important factors in the intelligibility of concurrent speech streams (see Section 3.3.3). Television use cases introduce a different set of requirements and challenges, however, regarding acceptable workload levels, interaction, and use context. Despite not being available in commercial products, some researchers have investigated and developed auditory displays for computer access relying on concurrent speech streams.

#### **AudioStreamer and Audio Hallway**

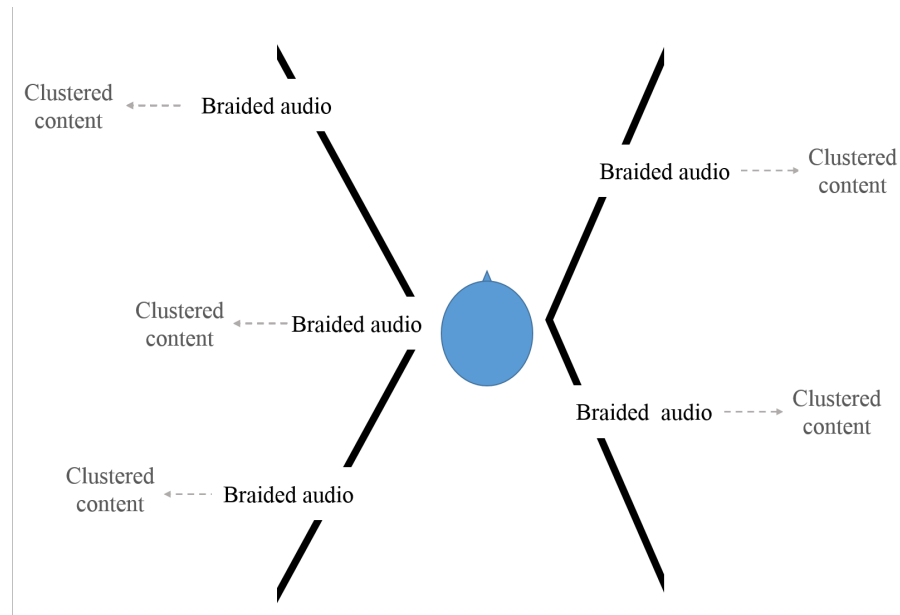
Schmandt & Mullins (1995) proposed *AudioStreamer*, which was one of the first auditory displays that attempted to utilise peoples' ability to selectively attend to a desired stream of speech in the presence of other talkers. The system presented three concurrent spatialised speech streams, which were binaurally spatialised to  $0^\circ$  and  $\pm 60^\circ$  in order to make use of the cocktail party effect. It is notable that the authors chose not to separate the current items maximally (i.e.,  $0^\circ$  and  $\pm 90^\circ$ ). These positions were chosen "to be large enough to

allow easy perceptual segregation of the sources, but still limit the time it takes to switch from one to the other, which is proportional to angle” (Schmandt & Mullins, 1995, p. 218). Mullins (1996) indicates that this decision was based on Rhodes’s (1987) findings of increased reaction times for increased angular separation in non-speech localisation tasks. More recent experiments, however, have found no significant differences from increased switching angles, though as the angular separation from an attended location increases, response times increase (Mondor & Zatorre, 1995). In addition to spatial separation, the display also used different talkers for each of the streams so as to exploit acoustic variations and reduce informational and energetic masking between concurrent talkers.

Perhaps the most interesting feature of the *AudioStreamer* display was the attempt to adapt to a user’s interest in one of the presented streams by analysing head movements. If the user turned their head towards a particular source, its level was temporarily increased, then exponentially decreased over time to return to the original level (Schmandt & Mullins, 1995). If the user wished to isolate one stream they could repeatedly look toward the virtual source, in which case the other streams would be silenced. The system was also designed to momentarily draw attention to key points in other streams so as to avoid important sections being missed.

Mullins’ masters dissertation went into further detail about the development of the *AudioStreamer* display (Mullins, 1996). In it, Mullins states that participants were overwhelmed by three channels of simultaneous speech and therefore introduced five-second onset asynchronies between each stream. Such a large onset asynchrony vastly exceeds the length of a word and therefore is not comparable to the studies reviewed in the discussion of onset asynchrony in Section 3.3.3. Unfortunately, despite the development, no formalised experimentation was presented in either work, making it difficult to assess how effective the display was, either in terms of communicating the information or of the user experience it provided.

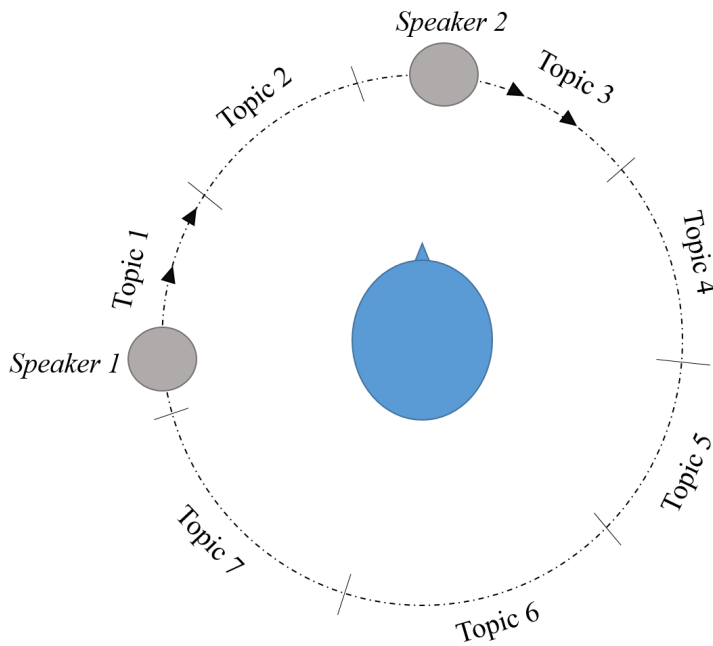
Schmandt (1998) proposed a second auditory display exploiting concurrent speech presentation called *Audio Hallway*. Similar to the *AudioStreamer*, it was intended to be used to allow the browsing of large collections of audio files. The *Audio Hallway* display provided the user with two levels of navigation; one ‘high-level’ allowing the navigation of groups of clustered content and a ‘low-level’ navigation of the individual audio files within a selected cluster. The high-level navigation was facilitated using the metaphor of a hallway in which doors were situated on either side leading to rooms filled with clustered content. The



**Figure 4.2:** Visual representation of the ‘Audio Hallway’ display. Adapted from Schmandt (1998, p. 167).

users travelled down the hallway using a head movement either forward or back and entered a door by tilting their head to the corresponding side. The doors were denoted by presenting all of the grouped items concurrently and automating the gain of each item so that individual items took it in turns to be the most prominent, a method which was termed as *braided audio*. The hallway was rendered binaurally, with three clusters audible simultaneously, such that the closest door was heard on one side with the next and previous doors on the other side perceived as in front or behind the listener respectively. Azimuthal distance between concurrent clusters was increased by creating a model where the hallway increased in width the further away it was from the listener’s position (see Figure 4.2). The cluster closest to the listener was presented more loudly than the other two sources to make it more prominent and therefore more easily attended. Despite these modifications users reportedly struggled with this display, which was interpreted by the author as an indication that combining multiple spatially separated sounds with listener position movement was not appropriate for auditory displays.

The low-level navigation in *Audio Hallway* was provided once the user had entered one of the rooms. Up to twenty items were presented on the azimuthal plane in the frontal hemisphere with up to four active at any one time. The horizontal location of the sources was distorted according to the orientation of the listeners head so that the spatial separation between the items was exaggerated. To emphasise this effect, the gain of items in front of the listener

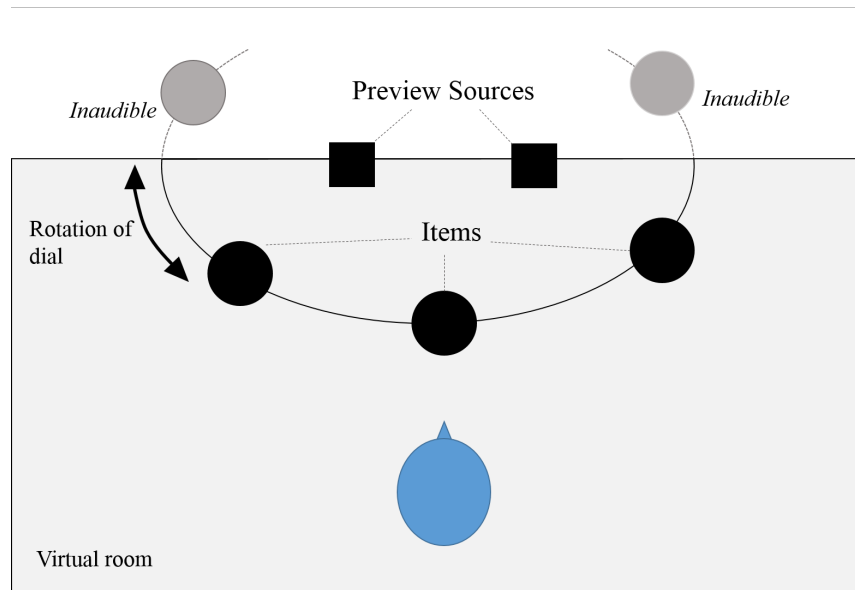


**Figure 4.3:** Visual representation of the ‘Dynamic Soundscape’ display. Adapted from Kobayashi & Schmandt (1997, p. 167).

was higher than those towards the sides. Although no formal user testing was described, Schmandt (1998) reported that users struggled less with the navigation and attributed this to the locations of the auditory items being more easily associated with the orientation of the listener’s head.

### Kobayashi and Schmandt’s Dynamic Soundscape

Kobayashi & Schmandt (1997) continued to research into using multiple concurrent streams for browsing speech audio with the *Dynamic Soundscape* system. The system positioned virtual sources, termed ‘speakers’, around the head on the azimuthal plane using binaural techniques. This distributed the content around the user’s head, which each of the speakers would play as they orbited the user (see Figure 4.3). The user controlled the interface with a touch pad, knob or through pointing gestures, which they used to activate a maximum of four speakers at any time. A head tracker system was also used and, like *AudioStreamer* (Mullins, 1996), head movements were analysed to assess user attention to specific speakers and alter their relative prominence through level manipulations. A continuously playing *audio cursor* served to indicate the position of the user control (hand, or point on touchpad), which was found to be especially useful for users for whom the binaural rendering was less effective.



**Figure 4.4:** Visual representation of the virtual dial display. Adapted from Frauenberger & Stockman (2006, p. 143).

The display highlighted issues with the loss of spatial resolution in memory, which meant users struggled to remember the precise locations of specific content (Kobayashi & Schmandt, 1997). This issue may have been exacerbated by the movement of multiple speakers and the concept of spatially distributed continuous information, which would be expected to impair spatial acuity compared with discrete points in a static interface. Unfortunately, though some user testing of the interface was performed, no comparisons were made between this and a traditional serial display with ‘fast-forward’ or ‘rewind’ functionality, making it hard to assess how beneficial this design would be.

### Frauenberger and Stockman’s virtual dial

Frauenberger & Stockman (2006) proposed a design using concurrent speech to navigate auditory menus using the idea of a virtual horizontal dial with items located around its perimeter. The display used a virtual room with the centre of the dial positioned outside so that a maximum of three items from the menu would be inside the room, and therefore audible, at any one time (Frauenberger, 2013). Two additional ‘preview sources’ were also audible if the selected item was a sub-menu (Frauenberger & Stockman, 2006) (see Figure 4.4). The user navigated the menu by rotating the virtual ring using a game pad dial until the desired item was directly in front. The display made use of different voice identities and talking styles (i.e., voiced or whispered) to reduce between-stream confusions.

The system was experimentally evaluated against a traditional screen-reader interface. Results indicated that performance was initially faster with the prototype interface, but performance significantly improved with the traditional screen-reader in the second trial. This, combined with a slight increase in task completion time with the prototype, led to the traditional interface becoming faster. Frauenberger & Stockman (2006) suggested that this phenomenon may have been due to the fatiguing effect of the constantly repeating audio, as participants commented on this being exhausting.

From a navigational speed perspective, the system's design seems unlikely to have been optimal. The prototype interface included some redundancy in the display as, following the initial display of three items, each subsequent display contained only one new item. This effectively reduced the display to a serial presentation with a reduced SNR. Furthermore, selection required the user to position the target item at the central location. This would have meant that a target detected at a lateral location would have had to be repositioned before it could be selected. While this reduced the amount of hardware required to make selections, it would have taken participants longer than if they had been able to select any of the audible options.

### **Clique**

Parente (2008) proposed an auditory display system using concurrent speech for computer-based GUI tasks, which was named *Clique*. The system was designed to use a collection of *views*, separating required information into different levels of relevance and similarity. Information that was most likely to be of interest to the current task was presented as part of the primary view. The preview provided information on items or tasks that were part of, or could become part of, the target task (e.g., a summary of the length of an email). These views were reminders of context (i.e., what application was in use), referred to as the overview. They could be accessed by the user at any point and therefore took the form of a background ambiance. A peripheral view was included to provide notifications regarding tasks completed by other tasks, such as the arrival of an email. A final view enabled the user to repeat the output of the other streams to compensate for the issues with memory; this was named the review.

One of *Clique*'s key features was that it was designed to separate the user experience from the underlying GUI, allowing consistent patterns of interaction to be deployed over different applications. The system received commands through keyboard shortcuts and allowed



functions such as searching to be conducted over all applications. The interface presented concurrent streams of information provided by ‘virtual assistants’, each tasked with providing specific information. The *content* and the *narrator* formed the *primary* view, with the content providing information such as the text in an email and the narrator producing sounds to echo user inputs. The *summary* acted as part of the preview, providing information on the number of emails in the inbox or the amount of time it would take to read a presentation. The *related assistant* provided some information for the preview as well as some information for the peripheral, on any state changes within the current task. The *unrelated assistant* acted as part of the peripheral view, providing information on other tasks or subtask state changes, such as an email arrival. The *environment* provided the context view in the form of atmospheric sounds and were presented to the listener without spatialisation.

Assistants were positioned at distinct points in space on the azimuthal plane using 3D audio. Different voices were assigned to the *assistants* to improve the user’s ability to distinguish concurrently presented content and a 200 ms onset asynchrony was included to assist with stream segregation. A mathematical proof was provided which demonstrated that concurrent presentation with an onset asynchrony would allow faster access to information than a serial interface does. In addition to the spoken content, the system used a combination of earcons, auditory icons and speech, depending on the nature of the content being expressed. States and actions were generally expressed using earcons, whilst auditory icons were reserved for identifying the type of subtask (i.e., list, table etc.).

The interface was assessed experimentally with both visually-impaired and normal-sighted participants in two separate trials. The assessment with visually-impaired participants comprised several distinct tasks. It explored participants’ ability to recall information from a target stream and unattended streams with different numbers of maskers. Performance was compared with a commercial screen reader (JAWS) in terms of finding specified target items, learnability, and multi-tasking performance.

In the comparisons of performance with the prototype display with different numbers of competing streams, the results showed that the participants had significantly higher success rates for the target speech over the secondary streams. For the target information, the only significant difference was between the two and three concurrent streams, while the non-target results showed a significant difference between all the numbers of streams. When the interface was compared with JAWS, performance at finding specified target items appeared to depend on the capabilities of the system being represented, with Clique allowing for significantly

more successful selections only where there was no search capability provided in the JAWS version of the interface. The learnability assessment results indicated that Clique led to more correct descriptions of how to complete tasks after a short training phase than JAWS did, which was attributed to the fewer commands required by the user to perform tasks. A multitasking assessment found that participants completed significantly more tasks with JAWS than Clique. Parente attributed this result to the marking system used, whereby the partial completions were not included. Assessment by normal-sighted participants found that if users started tasks using Clique they needed less time to complete them with the GUI later. The total time spent interacting with the interface was, however, substantially greater. It was also found that users preferred a simplified version of Clique which imposed a reduced workload.

Despite the apparent success of the Clique system in outscoring the JAWS interface in the majority experiments presented, it is unclear whether the final system is truly optimal. Though the use of asynchronous onsets in conjunction with gender and pitch differences is undoubtedly advantageous, justification is not provided for the final combination of parameters, which makes it hard to apply the findings to the design of future systems. Furthermore, interpretation of task durations was complicated due to the different interaction capabilities of the different displays (i.e., availability of search functions) and therefore it is difficult to determine how much advantage was provided through the use of concurrent speech.

### **VoiceScapes**

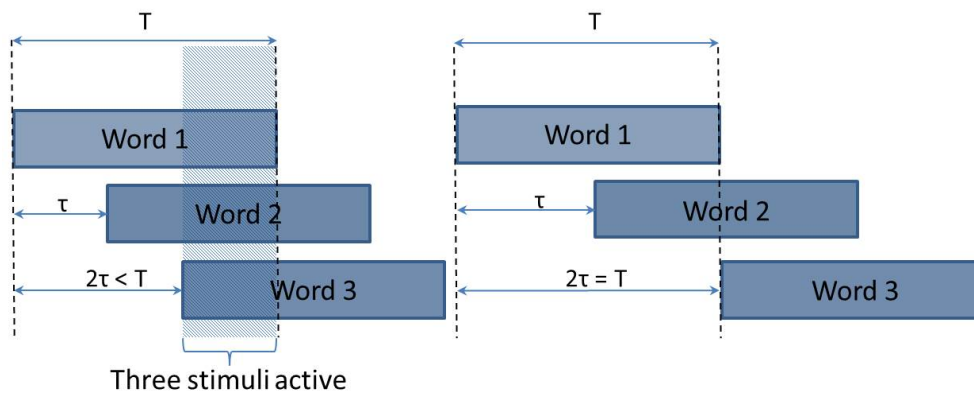
Werner *et al.* (2015) proposed a menu display using concurrent streams of speech in which between three and seven talkers were presented concurrently. Each source was presented from a unique spatial location and each was associated with different voices, which were arranged so that adjacent talkers were of different sexes. Interestingly, spoken items within the menus were looped. As words were of different durations, this meant the relative phase of the items changed during the presentation. A pilot study compared serial presentation of normal and of compressed versions with two different concurrent presentations. Although it is not entirely clear in the original paper, it is believed that one of the displays increased the number of talkers present (presenting the first three items before adding two more and then a further two), allowing the users to hear the display with fewer active talkers before more were added. The other concurrent presentation did the opposite, starting with seven concurrent items

and finishing with three. Results indicated that normal speed serial speech presentation was faster and easier to use than the concurrent speech approaches.

### **vCocktail**

The *vCocktail* system by Ikei *et al.* (2006) made use of onset asynchrony between overlapping successive spoken menu items, which they referred to as ‘multiplexed speech’. The authors conducted a series of experiments to determine the optimal configuration of the display. The first experiment investigated the localisation accuracy achieved using the system. This entailed randomly selecting one of the 40 words and 36 directions, presenting it to the participant over headphones and waiting for a response indicating its location on the user-controlled GUI. The results showed front-back confusion to be a considerable issue. After excluding these errors, however, practically all of the localisation results exhibited a mean error of  $\approx 20^\circ$ . A second experiment presented a display of between two and four concurrent words and then tasked the user with identifying a word that had been present from a list of on-screen words. The onset asynchrony was systematically varied between 0 and 500 ms in 100 ms steps. The presentation was either diotic or spatialised in the frontal hemisphere with equally spaced angular intervals ( $180^\circ$ ,  $90^\circ$  or  $60^\circ$ ) for two, three and four item presentations. The results showed that performance was improved by the introduction of onset asynchrony and by the spatialisation. No further significant gain was observed by increasing the onset asynchrony past 300 ms, which was approximately half the word duration. It was found that the spatialisation gave most advantage over diotic presentation when the onset asynchrony was low.

In the final experiment, the authors spatialised speech sources using onset intervals which ranged between zero and 500 ms. These were assessed using between two and four sources, different source orderings (i.e., from left to right or alternating between sources from either hemisphere) and either with or without a linear increase in attenuation applied over the course of each word. It was found that with three or more voices, high accuracy ( $\geq 99.7\%$ ) could be achieved with onset delays of 200 ms. This increased to 300 or 400 ms if adjacent sources were used and no attenuation applied in the three- or four-source conditions, respectively. As noted by Ikei *et al.* (2006), the optimal asynchrony without attenuation (300 ms) is about half of the duration of the stimuli (530 - 600 ms). This is important when considering the playback of more than two voices, because when the onset asynchrony is less than half the duration of the stimuli, all three items overlap, but when the onset asynchrony is half the



**Figure 4.5:** The reduction in the number of concurrent sources when an onset asynchrony of 50% or greater is used (adapted from Ikei et al. (2006, p. 192)).

duration or greater, a maximum of two items are presented at any one time (Figure 4.5). This simplifies the task in terms of both attentional load and SNR. If the optimal asynchrony were purely due to this effect, a similar behaviour would not be expected within the two-talker condition. The two-talker results appear to indicate improving performance with increasing onset asynchrony, but ceiling effects in the participants' performances make it difficult to ascertain the strength of this effect.

The study explains the need for concurrency to allow faster information flow and consequentially quicker browsing. Interestingly, as no investigation was performed into the effect that the conditions had on the response times of the participants, it is unclear whether the system offered any advantage over a traditional serial speech display.

Saito *et al.* (2010) developed the VCocktail+ display which used four concurrent talkers with a 300 ms onset asynchrony, cross-ordering and attenuation. Users interacted with the display using head gestures. A study compared user performance when switching between solving maths problems and navigating a menu within a music player when using VCocktail+ and the standard visual interface on the same screen using a mouse. The results show no significant differences in the time taken or the error rate between the two conditions, though this may be due to low number of participants (6). This finding appears positive for the VCocktail+ system. As with the original VCocktail work (Ikei *et al.*, 2006), however, it is not clear how this would compare to a serial presentation in terms of speed or other experiential aspects (e.g., workload). Furthermore, as the study was not solely focussed on the display component, the visual condition involved switching between windows and a different input mechanism. It is unclear to what degree these differences affected the results.

### **Vazquez-Alvarez and Brewster's Eye's-Free multi-tasking**

Most of the use cases discussed so far have been concerned with presenting streams of information that form parts of the same task (e.g., items in a menu or elements of an application). Another potential use is where the user is engaged in multiple concurrent tasks. Vazquez-Alvarez & Brewster (2010, 2011) explored how auditory menu navigation may be performed while simultaneously listening to a spoken podcast or music. They experimentally compared the impact of pausing the podcast during menu navigations, with concurrent presentations in which: the sources were collocated at the front, the menu was presented from the front while the podcast was 90° to the right, and the podcast moved from the centre to the right-hand location during menu navigations that were presented from the centre (termed 'spatial minimization'). With the podcast, participants were asked to try to attend to both streams (referred to as divided attention), while with the music, participants were given instructions to attend to the menu only (referred to as selective attention).

The results presented by Vazquez-Alvarez & Brewster (2010, 2011) indicated that for the divided attention task, interrupting the podcast resulted in lower workloads and shorter completion times, and was generally preferred to concurrent presentations. The different spatial presentations in the concurrent conditions appeared to have little impact on any of the measured factors. In the selective attention task with music, participants' workloads and completion times were lower, and there was less of a preference towards interrupting the music.

### **Text-To-Speeches**

Guerreiro (2013) suggested the use of concurrent speech presentations for speeding up the scanning of web-pages for users of screen readers. Following this, Guerreiro & Gonçalves (2014, 2016) created the text-to-speeches system which presented concurrent speech streams. Binaural processing positioned the speech sources in the frontal hemisphere, maximally separated on the azimuthal plane. The effects of the number of talkers (2, 3, 4) and the differences (same, small and large) between voices in terms of pitch and vocal-tract length (manipulated together) were explored experimentally with visually impaired participants (Guerreiro & Gonçalves, 2014) and the results were compared with normal sighted participants (Guerreiro & Gonçalves, 2016). Each source presented a news story as speech. Participants were given a hint consisting of words from the start of the story, and they were asked to listen to that stream. Performance was measured by their

ability to identify which source the target had been presented from, the number of elements the participants recalled, and their responses to questions about the story's content. Results indicated that identification was best with either two or three concurrent talkers, though three-talker presentation caused a slight drop in performance. Interestingly, the voice modifications were not found to make a difference to performance, but were favoured by participants. It was also noted that participants often chose to identify the source by its location. The results also showed that there was no difference between visually impaired and normal-sighted participant groups, which was taken to suggest that concurrent speech could be used by all, regardless of their visual capabilities (Guerreiro & Gonçalves, 2016). Subjective ratings from both groups indicated that when there were two talkers, participants were comfortable and able to understand the target stream. Increasing the number of talkers had a negative effect on these ratings, leading to neutral ratings when three talkers were present and participants indicating that listening to four concurrent voices was difficult and not comfortable (Guerreiro & Gonçalves, 2016).

## 4.4 Summary

Audition clearly has great potential as a modality for the presentation of information in HCI. Auditory display encompasses a large variety of design approaches, many of which have been discussed within this chapter.

Serial speech interfaces are effective and with the availability of TTS technology, they are able to deal with a huge range of scenarios. This has led to their widespread use in commercial auditory displays. As a result of this, they are considered to be the default means of representing information through sound. Restrictions in terms of presentation speed are, however, likely to be problematic when a large amount of information needs to be represented. Furthermore, in scenarios with more than one source of content (e.g., synchronous companion experiences), imposing serial speech presentation may have a negative impact on the user's experience. Serial speech may, therefore, not always be the best approach.

The use of non-speech within computer interfaces, as an alternative to speech, has clear advantages in terms of the amount of time needed to present an idea. Of these approaches, spearcons appear to be particularly useful in HCI scenarios. In the use cases under consideration within this work, information is likely to be regularly changed (e.g., in the case of EPGs) or the information that needs communicating will be semantically complex.

This is problematic for all of the non-speech representations and so it is felt that a purely non-speech approach is unlikely to be optimal.

Studies which have proposed the use of concurrent auditory presentations have been reviewed and discussed. These appear to suggest that users do have a limited capacity for taking advantage of concurrent speech presentations. These displays are attractive, as they avoid the requirement of prior knowledge on the part of the user and, through TTS technology, they are able to represent any information that may be conveyed textually.

In scenarios in which content is displayed to the user, who then elects to attend to a particular stream (e.g., Guerreiro & Gonçalves, 2014), it appears indisputable that concurrency has the potential to reduce presentation times. When users are expected to interact with concurrent speech displays, however, a reduction in task time is not a foregone conclusion. Users may take longer to respond due to the additional processing required to disentangle the contents of the display, or they may be more likely to make mistakes due to reduced intelligibility. Much of the work reviewed in this chapter has been focussed on representing menus, but there has yet to be conclusive proof of an increase in navigational speed due to concurrency. In fact, some work indicates that the opposite is true (Werner *et al.*, 2015). In these scenarios, the use of onset asynchrony, alongside pitch and spatial separation of speech streams, shows promise for reducing navigation times. Due to the inherent trade-off between high response accuracy (Ikei *et al.*, 2006) and low overall presentation time, onset asynchrony is a factor which needs further consideration regarding its impact on overall navigation times. In addition to this, it is still unclear whether a designer who is creating a multi-talker display should configure it based on the amount of onset asynchrony or overlap. The effects of overlap and onset asynchrony in multi-talker menu displays with a particular focus on navigation time is, therefore, an area that requires further investigation and is considered further in Chapter 5. Navigational speed is not the only relevant factor in non-visual television experiences. Other use cases, such as those described in Guerreiro & Gonçalves (2014, 2016) and Vazquez-Alvarez & Brewster (2011) allow users to choose between different concurrent streams of informational content. A particularly interesting instance of this would be orchestrated synchronous companion experiences, which are identified as one of the experiential elements of interest in Chapter 2. With these experiences, users have a stream of main programme content and an additional stream of secondary programme content, meaning they must attempt to attend selectively to one of the streams, or divide their attentional resources between the two streams. This use case differs considerably from those discussed in this chapter due to

the multi-modal nature of television content and the complexity of the relationship between the two streams of content. This use case, therefore, raises interesting questions about how much concurrency should be present and the impact of an additional auditory stream on the experience of watching a television programme. This is explored further in Chapters 6 and 7.



## Chapter 5

# Menu navigation

### 5.1 Introduction

In Chapter 2, menus are identified as a central part of modern television user interfaces, and an element of the experience that is likely to remain important in future television systems. Some menu systems are likely to be used every time someone wishes to choose a new programme, so their design can have a large impact on a user's experience of the system. It is important that users who are unable to visually attend to an on-screen menu have access to a non-visual alternative so that they may navigate and select from available programmes, choose applications to launch on their connected television and configure their device to suit their requirements.

The design of non-visual menus has received attention from many researchers within the HCI and auditory display communities and several approaches were discussed in Chapter 4. Most of these had a focus on providing access to PCs for those with visual impairments or who are blind. As the boundaries between television, computer and mobile phone disappear, it is apparent that many findings are transferable between these fields. This is likely to be particularly true for menu navigation. There are several factors with television menu navigation systems that require special consideration.

- **Content is not static**—Television menus are likely to change on a regular basis in terms of their contents. EPGs, for example, are different every day and online services regularly update the available content.
- **Number of items can be extremely large**—Particularly in the case of menus

containing programme content, there can be a vast collection of available options.

- **Multiple users**—Connected televisions are likely to remain multi-user devices and so must be usable by different individuals interchangeably.
- **Browsing and searching**—Users will often not have a strongly defined target in mind when searching for programmes (Elsweiler *et al.*, 2010). The systems must, therefore, cater for users with a definite target (e.g., *the next episode of ‘Sherlock’*) or only a vague criterion (e.g., *the least disagreeable live option*).
- **Relative amount of use**—some elements of the interface will see heavy use (e.g., programme selection menus), while others are likely to be used only very occasionally (e.g., utility menus).

From the discussion of different auditory display techniques, it is clear that a non-speech approach is inappropriate for many television uses. Utility menus are likely to have constant structure and contents, a relatively small number of options and are unlikely to be used on a regular basis. Whilst the small number of options is an advantage, the learning of semantic links is more problematic due to the user encountering them only occasionally. This could suggest that auditory icons with metaphorical or nomic representations would be a suitable alternative, as they have a more obvious link to the content they represent. Many of the items found in these menus, however, have no obvious metaphorical mapping (e.g., ‘picture settings’). Spearcons are an alternative approach, but the reliance on recognition rather than intelligibility means that users are required to have a good idea of the names of objects in the menu being navigated. This suggests that a speech-based approach is required. The additional time requirement of serial speech in these contexts is unlikely to be excessively problematic due to the comparatively small number of items to be represented and their infrequent use.

Conversely, programme selection menus are likely to see regular use, but will comprise a large number of options. Speed of use will be a much more significant factor in users’ experiences. These are problematic for non-speech representations due to the number of items involved and the frequent addition of new items. It is possible that navigation through a hierarchical structuring of content (e.g., by genre or format) may benefit from non-speech enhancements. In many cases, a reasonably large number of items may still be present at the lowest node of the menu (e.g., television programmes → comedy → sit-coms). At this point, non-speech representations are inappropriate and serial speech interfaces may be frustratingly slow.

Concurrent spoken menus offer an alternative to serial speech for both types of menu. As new representations do not have to be learnt, their usability is unaffected by changing content and by long intervals between uses. As concurrent displays allow for information to be displayed more quickly than a serial equivalent, they have the potential to facilitate faster, more efficient navigation. Following the work reviewed in the previous chapter, however, it is unclear if speed improvements are actually possible within these displays and whether it is onset asynchrony or the proportion by which words overlap that should be controlled. This chapter presents the design of a display based on these principles and describes an experimental investigation into how manipulating onset asynchrony and overlap affect navigation speed, accuracy, and workload.

## 5.2 Display design

In any investigation of the use of concurrency in auditory displays, it is clear that the precise design of the system is a crucial factor. An inefficient design may negate any potential benefits in terms of navigation speed, or result in unnecessary workloads. This section considers different factors of the display's design with this in mind.

### 5.2.1 Number of talkers

The greater the number of concurrent talkers, the larger the potential saving in terms of navigation times. On the other hand, when using a concurrent auditory presentation of speech, users find it harder to identify and detect accurately the displayed words as the number of concurrent speech streams is increased (Shafiro & Gygi, 2007; Nelson *et al.*, 1999). This issue is moderated when onset asynchrony is introduced into the display and, by using this approach, Ikei *et al.* (2006) found that very high accuracy can be maintained for greater numbers of talkers.

Benefits of dividing the options into shorter groups extends beyond considerations of intelligibility. Even if it were possible for a user to hear all of the items, as the list increases in length it becomes harder for users to remember all of the options and the interface used in the selection becomes more complex. Grouping spoken lists by adding pauses improves recall performance (Ryan, 1969). Some grouping is therefore likely to be beneficial if users are browsing and wish to choose the most appropriate of the available options.

Three concurrent talkers will be used in this design. It is feasible that larger numbers of talkers

could also prove beneficial. This configuration represents a compromise between complexity of the display, potential navigational speed advantage and consideration for spatial separation, discussed in the following section.

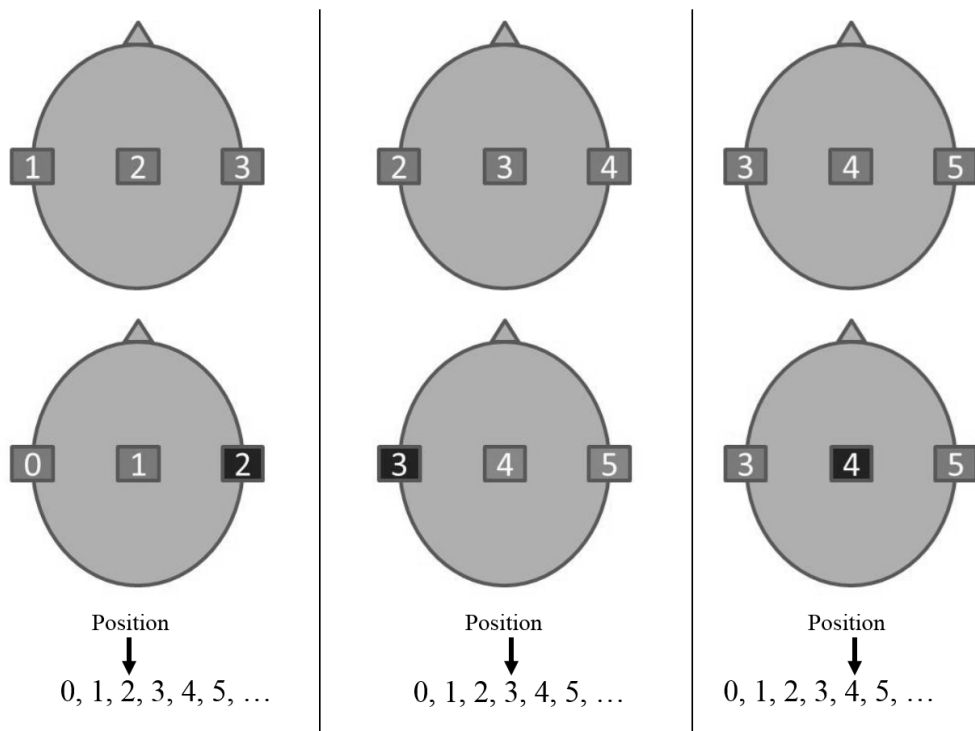
### 5.2.2 Spatial configuration

Many previous auditory displays using concurrent speech have made use of binaural processing to spatialise the audio sources (e.g., Frauenberger & Stockman, 2006; Ikei *et al.*, 2006). It is not clear that this would be beneficial when there are only two or three talkers. Consider a binaural presentation of two talkers rendered with maximum lateral separation on the azimuthal plane. Both ipsilateral and contralateral ears receive a mixture of target and masker signals at slightly different levels, and the phase relationships between the signals differ in each ear. If a dichotic presentation were used each ear would only receive signals from one source. The dichotic presentation therefore provides a greater SNR in the ipsilateral ear, reducing the amount of energetic masking at the cochlea. A similar principle holds in a three-talker scenario with maximal azimuthal separation in the frontal hemisphere. Again, with binaural processing each ear receives a combination of signals from all sources. If the three sources are presented with one source restricted to each ear, and the third presented equally to both, each ear only receives two signals, the ipsilateral and central sources. While energetic masking will occur in this instance, it would be expected to be less, as it would involve two signals rather than three. Furthermore, relying on intensity panning still results in the signals appearing to come from different lateral locations, which would be expected to lead to spatial release from informational masking. With larger numbers of talkers, however, the number of sources at intermediate positions increases and the differences between source locations will be smaller. In this situation, the use of intensity panning is likely to become less effective for distinguishing sources than binaural rendering.

Furthermore, binaural spatialisation introduces perceptual factors which tend to vary between participants (e.g., externalisation and front-back confusions), whereas intensity-panning does not. From an experimental perspective, intensity panning is considered to be a more robust approach for providing consistent experiences across participants.

### 5.2.3 Interaction design

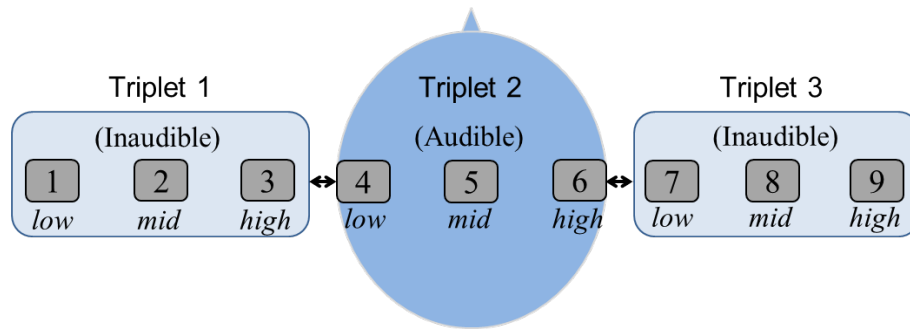
In considering designs for the user interaction within this work, designs in which the user actively moves through the list of options were considered. Perhaps the simplest interaction



**Figure 5.1:** Illustration of the sliding window (upper) and segmented (lower) interaction methods for different positions in the list. In position 3 (central column) the sliding window approach repeats items 2 and 3, which were presented at position 2. Conversely, in the segmented model, at position 3 all three options differ from those presented in position 2. In the sliding window model, the item in the centre location can be selected. In the segmented model, the darker box indicates the item that can be selected.

design is to play the user a spoken list of grouped items and allow them to respond when they hear the item they are looking for. This reduces the number of interactions that are required from the user. The pacing of such a display is critical. If the pacing of grouped items is too fast, users do not have enough time to resolve if the desired target is present and could feel that they are getting left behind. Conversely, a pacing that is too slow would negate any advantage of the display and lead to further frustration. An additional complication is that the definition of an acceptable pacing could be dependent on the individual, their context, the combination of words, and the onset asynchrony. A considerable amount of further work would be required to determine if there is a universal optimum for the pacing of these displays. This is beyond the scope of this project.

During the design, several different models of interaction were considered. Like the design proposed by Frauenberger & Stockman (2006), the first concept allows the user to scroll through the list, hearing items move from left to right as they do so. Frauenberger & Stockman (2006) referred to this as a ‘sliding window’ approach. In this model, a user



**Figure 5.2:** *Illustration of display design concept (italicised writing refers to stimuli pitch).*

must interact with the system to position the desired item in the centre location to be able to select it. The second ‘segmented’ approach presents the items in groups of three. The items remain in the same locations until the user moves on to the next group of three (Figure 5.1). Within this model the user may select one of the three currently audible options. To represent which item is selected, a short tone is played from that location or a differently pitched voice is used for the selected item.

Within these models, in which only one item can be selected at any moment (i.e., either the item in the central location or the one selected with the auditory cursor), the additional movement associated with the sliding window is not considered to be beneficial. On every interaction, all sources are repositioned, which means that a user must re-evaluate the composition of the auditory scene. Furthermore, as discussed in Section 4.3.3, this model becomes equivalent to serial navigation after the first presentation.

Forcing a user to select one of the audible items through manipulating the display is implicitly inefficient. After the user has identified the presence of the desired item, they must manipulate the display to reposition the item or the cursor so that it can be selected. By incorporating different commands to allow the user to select any of the items within the audible triplet, the user is able to react more quickly to a displayed target. Our system was therefore designed to use a segmented presentation where the user could select any of the audible options. This has important implications. It is not possible to make assumptions about which of the concurrent sources the user is likely to be trying to attend to and treatments that would detrimentally impact on the audibility of one source to improve the audibility of another should be avoided. The sources within a triplet were presented from maximally separated lateral positions using intensity panning, such that one source would appear on either side of the head through being presented to only the ipsilateral ear, whilst the third source was presented at the same level in both channels so as to appear in the centre of the head. The use of a symmetric

spatial configuration does lead to the possibility of a bias towards some sources as a result of the spatial asymmetries in human performance (e.g., Kimura, 1961a; Bolia *et al.*, 2001; Sætrevik, 2012). Symmetrical presentations have, however, been associated with improved performance in CRM tasks (Bolia *et al.*, 2001).

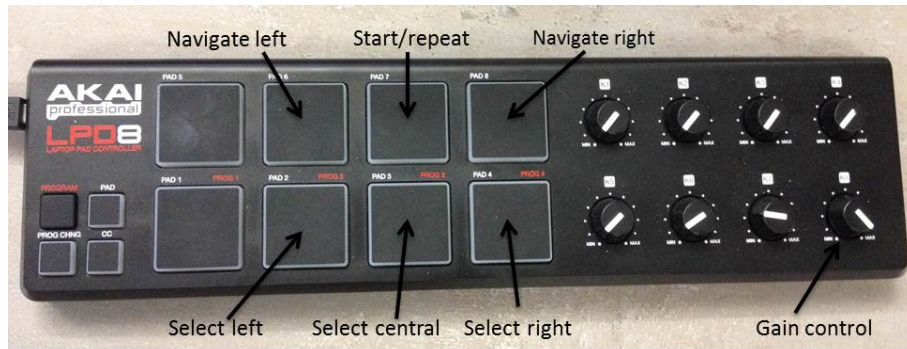
Within a triplet, sources were also distinguished by a pitch difference such that the stimulus on the left was lowest and the stimulus on the right was highest (See Figure 5.2). The order of presentation was kept constant and ran from left to right to correspond with normal reading direction. The use of three talkers means that the binary distinction of the talkers' sexes is not enough to identify each of the streams. Furthermore, using one talker who is of a different sex to the other voices may distract from the other two presented items (Brungart *et al.*, 2001). Other studies have found advantages in modifying the talker's apparent sex through vocal tract length modification (Darwin *et al.*, 2003), this was not attempted in this study to avoid making the voices sound excessively unnatural.

It is possible that a user may choose to interact with menus on a multitude of different devices which provide television experiences. Different devices will require users to control the display in different ways and may rely on different input modalities. As this project is focused primarily on display design, as opposed to the efficacy of different types of interactions, simplicity of use and implementation were prioritised over the creation of a real-world control interface. Furthermore, the focus on navigation time meant that it was important to minimise other factors that could contribute to inconsistent timing data (e.g., network communication delays). The system was controlled using five of the drum pads and one rotary dial on a universal serial bus (USB) musical instrument digital interface (MIDI) controller. Two rows of pads on the device were assigned to different functions. Playback controls (e.g., start playback, repeat triplet and navigation between triplets) were provided using three pads on the top row and selection of the three items in the currently selected triplet was performed with three pads on the bottom row (Figure 5.3).

To signal to the user when they had reached the end of the list, a short sine tone was played. This tone came was panned to come from the left or right channel when the user reached the left or right-most extremities of the list, respectively.

#### 5.2.4 Implementation

The prototype system was developed in Pure Data (Pd-extended 0.43.4) (Puredata, n.d.) and run on Ubuntu 12.04 LTS with a low latency kernel. To reduce computational load



*Figure 5.3: Control and function mappings used for the prototype display.*

and the possibility of error, the triplets were processed in advance, as discussed in Section 5.4.1. Therefore, the patch was mainly responsible for receiving the MIDI messages from the controller, handling trial configurations, file playback and recording responses.

## 5.3 Pilot Study

The pilot study was intended to investigate the effects of different amounts of onset asynchrony on the speed at which menu navigation could be performed and to compare this to performance with a serial display. One of the concerns with a concurrent auditory interface is how demanding the user would find it to use. Given the use case of menu navigation on television systems, users are unlikely to be willing to use the display if it is perceived as requiring an excessive amount of effort to use. It was therefore decided to assess the subjective impressions of the displays in addition to the performance measures.

### 5.3.1 Methodology

Eight participants were recruited from the BBC Future Media department. Potential participants were asked not to volunteer if they did not consider themselves to have normal hearing, but no audiometric testing was performed.

### Stimuli

Many studies on multi-talker displays have made use of the CRM corpus developed by Bolia *et al.* (2000) (e.g., Brungart, 2001; McAnally *et al.*, 2002). This methodology was not used here for several reasons. Firstly, the use of call signs has an inherently military feel and seems somewhat out of place within the use case. Furthermore, the call signs have



been developed specifically to be easily identifiable from each other under adverse acoustic conditions and can have very different phonetic structure. In fact, the start of each call sign completely distinguishes it from all the others. This suggests that the listener need only attend to this element of the mixture in order to identify which of the call signs are present. These factors may mean that the detection of a specified call sign performance would be higher than when less favourable combinations of words are used.

For the pilot, it was decided to use a collection of consonant-vowel-consonant (CVC) words that comprised two sets of eight words — one for training and one for the experiment. The experimental set consisted of four pairs of words with the same final vowel-consonant (VC) pair. Each pair shared the starting consonants with one other pair within the set (i.e., yawn, torn, youth, tooth, peace, lease, perch, lurch). The idea behind this was to control the differences between the words and, in particular, the part of the word which the participant would have to hear in order to accurately detect a target. The second set, which were for use in the training, were selected to have different phonemes from the experimental set and comprised: badge, wedge, good, shed, mash, fish, rang, herb. It was initially intended that this study would be followed by a second study looking at bisyllabic stimuli. Two additional sets of consonant-vowel-consonant-vowel (CVCV) were therefore recorded and processed alongside these stimuli.

The words were all recorded being spoken by one male talker at a sample rate of 48 kHz and 32 bits per sample. The words were processed in Praat (Boersma & Weenink, n.d.) to be a constant length of 530 ms and were resynthesised to have a constant pitch for when the word was positioned in the centre of a triplet. The pitches of the words on the right or the left in a triplet were adjusted approximately to plus or minus one ERB (Glasberg & Moore, 1990), respectively, compared to the pitch of the centre word. While other authors have reported benefits from much larger pitch differences (Darwin *et al.*, 2003), a comparatively small pitch difference was used here, as there was concern that more extreme modification would have jeopardised intelligibility and may have led to users becoming distracted by the unnatural character of the voices.

The individual stimuli were adjusted to have equal root mean square (RMS) values and then combined into triplets using MATLAB to provide onset asynchronies of 0, 80, 180, 280 and 380 ms. These asynchronies were chosen so as to have a condition near the centre of the word and two conditions occurring at different parts of the articulation for the bisyllabic stimuli. Stimuli were panned to their respective positions using the constant power pan law

(Reveillon cited in Roads (1996, p.460)) to maintain an equal power for each of the locations. This meant that the left- and right-most stimuli were presented with a gain of 1 in their respective channels, while the central source was presented with a gain of  $\sqrt{0.5}$  in both channels. For the serial condition, the individual normalised stimuli were exported. All of the final stimuli were exported as 44.1 kHz, 16-bit WAV files.

## Procedure

The independent variable was the presentation mode of the menu items, which was either concurrent with onset asynchronies of 0, 80, 180, 280 or 380, or serial. For the serial condition, participants navigated through the list using the same interface and controls, with the exception of the selection buttons. As only one word was presented at a time, this was output at the same level in both channels to appear central. In order to select a word, the user had to press the 'Select central' button on the interface. In this condition, the two other selection buttons did nothing.

The experiment was structured as a within-participants study, so that all participants experienced all of the conditions. Participants completed consent forms and then completed 6 blocks of training tasks using the training stimuli, one for each of the presentations. Following a break of approximately 20 minutes, participants completed 6 more blocks of trials, this time using the experimental stimuli.

Each block comprised individual trials in which participants were presented with a target word on a screen and then, when the user was ready, they pressed the 'start' pad, which immediately played the first triplet or word in the list. The user then navigated until they found the target, whereupon they could select it by pressing the appropriate pad. The target word remained displayed on the screen throughout the trial to reduce the possibility of a participant forgetting which word they were supposed to be looking for. In addition to this, the interface indicated whether the trial was currently on-going or not (see Figure 5.4). This took the simple form of a cross in a box, which disappeared when a trial was in progress. The main purpose of the indication was to inform participants about whether they had made a selection, or failed to start the next trial. During the training session, at the end of each trial, the identity of the selected word was displayed on the screen before the next target was presented.

Each block consisted of one trial at each of the target locations in a randomised order. In the experiment itself, participants were first played all of the stimuli they would be meeting.



**Figure 5.4:** *The visual display that presented participants with the target word and the current state of the interface. Screenshot has been cropped to remove excess whitespace*

The stimuli were presented in a pseudo-random order such that the same word at different pitches was never presented consecutively. After this, they performed another six blocks of trials, each containing one trial for each target position. The ordering of the presentation conditions in the training and main experiment were randomised so that conditions could not occur in the same block for more than two participants.

During the trials in the main experiment, participant interactions with the interface were logged to allow the calculation of task duration. While the measurement of task performance could be extracted from logging interactions with the interface, understanding the subjective demands of the different displays required an additional stage of data collection. For this, it was decided to use the NASA TLX subjective workload assessment (Hart & Staveland, 1988), which requires participants to rate their impression of temporal demand, mental demand, physical demand, effort, frustration and performance. These scales are then combined to provide an overall workload score. This method was chosen because it is relatively quick and simple to perform, and it has seen widespread use for interface design and evaluation since its inception (Hart, 2006).

Hart & Staveland (1988) proposed weighting the individual scores based on the participant's opinion of the relative contribution each factor had on the workload. As highlighted in Hart's review of the NASA TLX's usage (Hart, 2006), the weighting step is often omitted and experimental comparisons of weighted and unweighted versions have failed to reach a consensus on which method is superior. For this study, it was decided to omit the weighting step in the interest of minimising the experiment's duration and complexity. In order to gather TLX ratings for each of the presentation conditions, a computer-based version of the NASA TLX (Cao *et al.*, 2009) was undertaken by each participant at the end of each block

of trials.

After completing all of the experimental blocks, participants were asked questions on how they found the conditions in which the words overlapped, which of the displays they preferred, any strategies they developed during the concurrent presentations, and whether they had been aware of non-target words in the concurrent presentations. Also, as participants were recruited from the department in which the project had been completed and some presentation had been given on the project, participants were also asked about any experience they had of the stimuli or previous iterations of the interface.

### 5.3.2 Findings

The experiment was affected by a series of methodological and technical issues. As participants had only performed one trial at each target location for each onset asynchrony and incorrect trials could not be analysed, the data for navigational time was unbalanced. In addition to this, an issue with the prototype software meant that for each block a participant completed, the final trial's data was incomplete. Finally, a few trials were affected by procedural anomalies (e.g., a participant asking questions about the interface while completing a trial). This was particularly problematic because there was no way for the experimenter to flag specific trials as problematic at the time. Meaning that it was necessary to deduce which trial was affected. This lack of balance meant that the primary goal of the experiment, a statistical analysis of the navigation speed was not possible.

The navigation time data was noted to have a large positive skew with some exceptionally long task durations. It is believed that this skew was due to participants adopting different behaviours during the task. It was also noted that participants appeared to adopt different tactics regarding the use of the 'repeat' function. With some making regular use of the function, while others used it only very occasionally. As participants were able to navigate backwards and forwards in the list, there were instances where participants missed the target and navigated to the end of the list before returning. This non-uniformity of behaviour is also problematic from an experimental perspective as it could result in a multi-modal distribution for navigation time, and mean that participants' qualitative assessments are based on different experiences dependent on the method that they chose to use.

Unweighted workload scores were produced by the TLX software (Cao *et al.*, 2009), which took the mean of the ratings across the different factors for each participant in each block. Visual inspection of the results appeared to show that the workloads were higher for the 0

and 80 ms conditions. While a Friedman test on the results indicated a significant main effect ( $\chi^2(5) = 24.130, p < .001$ ), Bonferroni corrected pairwise sign tests failed to reach significance. Friedman tests of the TLX subscales also demonstrated significant main effects for frustration ( $\chi^2(5) = 17.895, p = .003$ ), effort ( $\chi^2(5) = 21.679, p = .001$ ), mental demand ( $\chi^2(5) = 21.415, p = .001$ ) and performance ( $\chi^2(5) = 22.860, p < .001$ ). Analyses of the other subscales did not indicate significant main effects ( $p > 0.05$ ). Again, *post hoc* pairwise Bonferroni sign tests on the subscales for which significant effects had been observed did not reveal any significant differences.

## 5.4 Experiment Proper

Following the pilot study, it was clear that some considerable methodological changes were necessary to gather robust and statistically analysable data. Furthermore, it was decided to expand the study to incorporate more than one word duration, to allow the effects of overlap and onset asynchrony to be explored. The experiment was therefore intended to answer the question: ‘how does onset asynchrony and overlap effect the navigation speed, workload and accuracy within concurrent spoken menus?’.

From considering the design of the pilot study, it was also decided to omit the serial condition from the main study. Due to the number of conditions involved in the redesigned experiment and the difference in interaction styles between the concurrent and serial conditions, it was felt that any comparison between the grouped concurrent display and the individual serial display should be performed in a smaller, more focussed experiment.

### 5.4.1 Methodology

Sixteen participants, who had not taken part in the pilot study, were recruited from amongst the BBC Future Media department staff and its visitors. Volunteers who reported hearing impairments were not included in the study, but no audiometric testing was performed on the participants. No attempt was made to recruit participants with vision disabilities, since the utility of a non-visual display is not solely restricted to people with visual impairments or who are blind.

During the experiment two participants experienced a fault in the software. For one of the participants this fault affected the experimental trials; therefore, this participant was excluded from the study and an additional participant was recruited to fill the space.

## Stimuli

On consideration of the stimuli following the pilot study, it became clear that the random combination of the words within a triplet led to variations in difficulty for target detection. Specifically, some triplets would comprise words with greater phonetic similarity to the target than others. For example, if the target *torn* were presented alongside the maskers *yawn* and *tooth*, it was likely to be considerably more challenging than if the maskers were *peace* and *lease*. For this reason, a set of stimuli with a more consistent phonetic difference was chosen for use in the main experiment.

Wordlists from the modified rhyme test (House *et al.*, 1965) were used as the source of the words for the experiment. The modified rhyme test wordlists consist of two sets, each of which is made up of 25 lists of 6 words. Stimuli were selected from the first set, within which each list of words shares the same first consonant-vowel pair but differs in the final consonant (e.g., *page*, *pale*), or in some cases has no final consonant (e.g., *ray*). For this experiment, 22 of the lists were chosen and for each of the lists one word was removed. Removals were mostly made because the pronunciations required for words to share common vowel sounds deviated from that which would be expected with Received Pronunciation (e.g., *pass*, *pat*), the lack of a final consonant, or because some words were deemed too unusual or inappropriate for the experiment.

The 110 words which remained were recorded spoken by a male talker, who was asked to enounce with minimal intonation and variation in word duration. Recordings were captured as 24-bit audio files with a sample rate of 44.1 kHz. Words were cropped to remove silences before and after they were spoken. Where possible, the crop was made at a zero crossing. In a few cases, however, no suitable zero crossing was available and the crop introduced a small discontinuity into the waveform. In such cases, the word was auditioned to ensure that no click could be heard. For some fricative consonants, low-level sounds at the start or end of the word were removed if they were considered not to contribute to the intelligibility of the word. The durations of the stimuli were found to vary quite considerably, ranging from 301 to 709 ms with an average of 486 ms.

The stimuli were manipulated to have the same constant pitches and durations (either 360 or 600 ms) in Praat (Boersma & Weenink, n.d.). These durations were chosen to ensure that no word would be stretched to more than twice, or shortened to less than half, of its original length. The pitch values of words in the centre of a triplet were adjusted to correspond to the average pitch of all of the stimuli. The pitches of the words on the right or the left in

a triplet were adjusted approximately to plus or minus one ERB (Glasberg & Moore, 1990), respectively, compared to the pitch of the centre word.

The stimuli were processed such that the words appeared to be approximately the same loudness in all onset asynchrony conditions. The onset asynchrony varies the amount of overlap between the words and this varies the overall loudness of the presentation. While it would have been possible to normalise the loudness of all presentations, this would have altered the levels of the words between presentation conditions. It was decided that it would be more consistent to preserve the variations in overall presentation loudness. To achieve this effect the individual stimuli were adjusted by ear to be of equal loudness.

The stimuli were combined into triplets and onset asynchronies were adjusted using MATLAB in a similar manner to the concurrent stimuli in the pilot study. A notable exception, however, was the removal of the RMS normalisation stage as a result of the stimuli having already been adjusted to equal loudness by ear. As in the pilot study, stimuli were panned using the constant power pan law (Reveillon cited in Roads (1996, p.460)).

Each triplet presented in the experiment met the conditions that all words had to be from the same list, with the same word length, and each word could only appear once within that triplet. This resulted in the creation of 10,560 triplets. The highest peak amplitude in the set of triplets was found and used to calculate the scaling factor necessary to bring this peak to an amplitude of (+/-) 0.9999 so as to maximise SNR whilst avoiding any clipping distortion. This scaling factor was applied to all of the stimuli to ensure their relative loudness was not altered. The triplets were then exported as 44.1 kHz, 16-bit WAV files.

### **Prototype modifications**

Following the observations from the pilot study, a number of functions were removed from the prototype system for the purposes of the experiment. This included the functions that allowed a user to return to a previous triplet or repeat the current triplet. While it is acknowledged that both of these features would need to be provided in a real-world implementation, it was deemed necessary to maintain consistency between participants' experiences and reduce the likelihood of skewed or multi-modal task duration data.

It is also notable that if correct selections are not possible without the repeats, this is an indication that the words are not sufficiently clear. Similarly, navigating past a word could have a sizable and undesirable impact on navigation times in real-world implementations

when the number of options is large. The removal of these functions from the interface in the experiment forces users to either take the most direct route to the target or make an error which can be easily identified and eliminated from the results before analysis of the navigation times is undertaken. It should be noted that these limitations mean that more errors are to be expected than if these functions had been left in.

As participants could only navigate in one direction, the sine tones which signalled that the user had reached either end of the list were removed from the prototype. Continuing to navigate beyond the final triplet was instead taken as an indication that the participant had not detected the target in any of the presented triplets and ended the trial.

### Procedure

The independent variables in the experiment are onset asynchrony and word length. The onset asynchrony has four possible durations [180, 280, 380, 480 ms] and there are two durations [360, 600 ms] for the word length. The onset asynchrony values were chosen to have conditions that were close to 50% of the two word lengths and to ensure that there were conditions in which the shorter stimuli were no longer overlapping. In order to minimise the duration of the experiment, it was desirable to limit the number of onset asynchrony conditions. Following their apparently poor performance in the pilot study and the results of Ikei *et al.* (2006) indicating an optimum value would be expected at higher asynchronies, it was decided to omit lower onset asynchronies such as the 0 and 80 ms conditions from the pilot study.

The experiment was structured as a within-subjects design. Experimental trials were split into sessions of fixed word length. Each session contained four blocks in which the onset asynchrony was kept constant. This structure was imposed on the trials so that NASA TLX subjective workload assessments (Hart & Staveland, 1988) could be performed on each of the word length/onset asynchrony combinations. On completion of each block, the computer-based version of the evaluation (Cao *et al.*, 2009) was undertaken by each participant.

The structure of the trials was very similar to that of the pilot. First, the target word was presented on a screen and then, in their own time, the user started the trial by pressing the 'start' pad. They then proceeded to navigate through the list until they found the target, which they could select using the selection pads. An additional complexity of this study was that in some trials the target word was not present in the list, in which case the correct



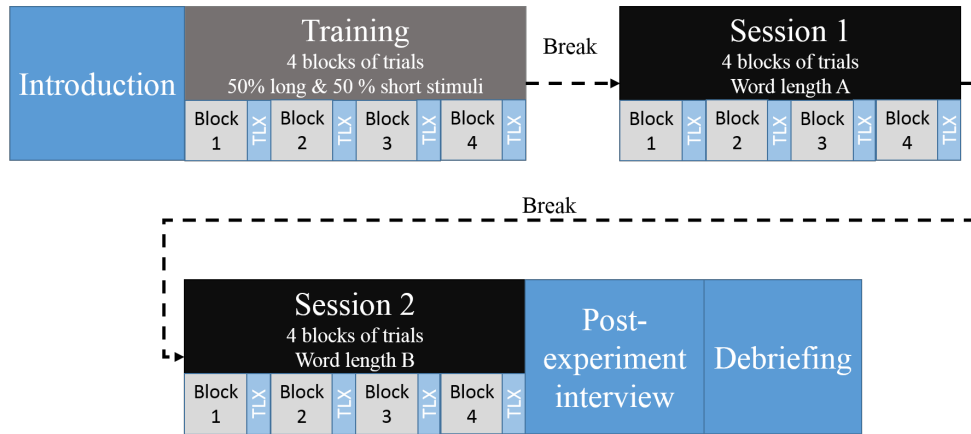
response was to navigate onwards from the final triplet. The visual interface was very similar to the one used in the pilot, in that the target was visible throughout the trial and the state of the current trial was indicated through the appearance of the cross in a box. Again, during the training session, feedback was given after each trial on the identity of the word the user had selected.

Each list in the experiment was nine words long, with the target words, if present, only presented once at one location. In trials in which the target was present the lists were constrained such that:

- each triplet had at least one stimulus that was not in the previous triplet
- all words in a triplet were different
- the target only appeared at the target location
- all non-target words had to appear twice and never in the same lateral location (e.g., if ‘page’ appeared in the central location in the first triplet of a list, it would appear in one other triplet but not in the central location).

These constraints were put in place to ensure that there was variation between the triplets in a list so as to avoid participants being able to rely on simply detecting that the triplet had changed. This led to 960 possible list combinations for each word at each location within the list. When the target was not included in the list, one item/word could appear three times, whilst all others would appear twice. In these instances, words would never appear in the same lateral positions. These criteria led to 576 possible lists for each word.

The experiment was split into three sessions with twenty-minute breaks between them to reduce the effects of any fatigue (see Figure 5.5). In the first (training) session, participants completed an informed consent form and then were introduced to the system. An initial playback level was set (participants had the ability to adjust this throughout the experiment) and they were given 40 practice tasks. The tasks consisted of 4 blocks of 10 trials, one block for each onset asynchrony. Each block consisted of all target locations and these were pseudorandomly allocated a word length condition such that 50% of each block was of each condition. Target locations were pseudorandomly ordered so that for each participant each target location appeared once for each trial number within a block. The participant completed a NASA TLX questionnaire for each of the blocks. Participants were given additional guidance when it appeared they had not fully understood how to use the setup or were unclear on how to respond to the TLX questions.



**Figure 5.5:** Diagram of the running order of the experiment. For half of the participants word length  $A$  and  $B$  corresponded to the short and long stimuli respectively. While for the other half of participants, the converse was true.

The second two sessions consisted of the experimental trials (limited to a maximum of 20 minutes each), with each session containing one of the word-length conditions. Half of the participants were presented with the short words first and the other half heard the long words first. To reduce the influence of ordering on the onset asynchrony conditions over the training and experimental sessions, three counterbalanced Latin squares were used to vary the orderings. For each instance of the Latin square the dummy values were substituted pseudorandomly for onset asynchrony conditions, such that no dummy value represented the same onset asynchrony condition twice. A row from each of the Latin squares was then used for each of the sessions. The order in which the Latin squares were used for the sessions was varied for every four participants to produce variations for all 16 participants.

Within each block, target location order was varied pseudorandomly, with the restriction that for each participant each target location could appear no more than twice at each trial index in the training and experimental sessions and not at the same trial index as in the previous session. Each trial's list was randomly chosen from the 22 possibilities such that the same list was never used in two consecutive trials. The target word was randomly chosen from the list. The experiment list was then randomly chosen from all possible lists for which the target was at the specified location.

To ensure that data points were captured for all target locations, trial accuracy information was output from Pure Data and read by a Python script. The script analysed the configuration of the original trials and generated new repeat trials when a participant failed to select the target. These repeat trials were then added to the list of trials being read by Pure Data. Repeat trials were reordered and modified so that they used a different one of

the 22 wordlists to both the preceding trial and the trial to be repeated. This ensured that the target identity and list were also different, so that a participant's previous exposure to the same target location would have minimal effect on their performance. Two additional dummy trials were added using target locations for which the user had already registered an accurate response. These served as a buffer zone in the event that the participant's response to the final trial in a block was incorrect. The Python script was used to edit the input to the Pure Data program while the participant was using it. The target location of the last output trial was used to decide where the repeated trials should be added. The repeats were added to the end of the original 10 trials until the last completed trial was beyond the eighth trial, at which point repeats were added after two trials. This ensured that no trials were altered after the user had already begun them and that a repeated trial was always separated from the original by at least two trials. For the repeats, the list was randomly selected from the 960 (or 576 if for a 'no-target' trial) possibilities, reducing the chances to a negligible level of a trial sharing the same list as another trial.

Following the experimental tasks, participants were asked for their opinions on how they found completing the tasks, what they thought of overlapping speech presentations compared to serial presentation, what strategies they used to detect the target, whether they were aware of the identity of words in the list other than the target, and how they felt the durations of the words contributed to their experiences of the tasks. For questions where participants were given grading options as part of the question (e.g., easier, made no difference, or made it more difficult), the ordering of options was reversed for half of the participants to control for recency bias. The researcher asked questions from the script (see Appendix B for the script that was used), but also asked for clarification on comments made by participants when the intended meaning was unclear.

The experiment was conducted in a user testing laboratory in the BBC Research and Development department. Participants sat on a sofa approximately 2.4 m in front of a television displaying the target word (see Figure 5.6). The controller was positioned in front of the participants on a coffee table.

### 5.4.2 Results

All statistical analysis was performed using SPSS. Further information on the statistical tests used can be found in (IBM, n.d.). During the running of the experiment, on four occasions it was clear that the participant made several attempts to navigate to the next triplet but had



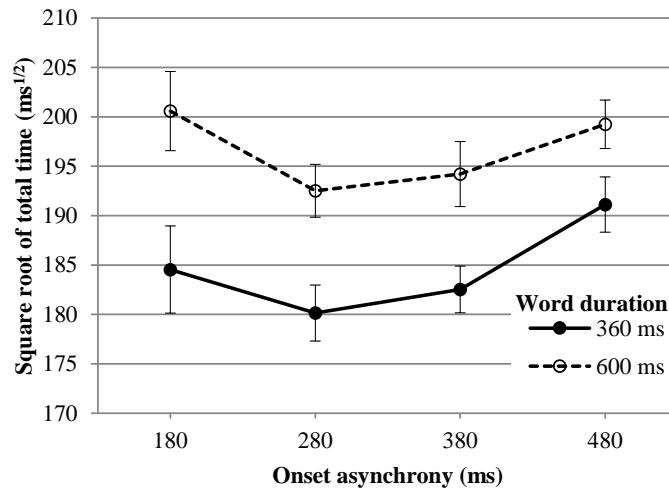
*Figure 5.6: The room setup used in the experiment*

not pressed the pad with sufficient force, causing a significant delay in their navigation time. The experimenter flagged these trials during the experiment and repeats were generated as if the trial had been incorrectly answered. Of the four affected trials, one had been a dummy trial. The original affected trials were removed from all subsequent analyses, with the data from the repeated trials being used in their place.

### **Total task duration**

The duration of a trial (i.e., the ‘task’) was taken as the time from the playback of the first triplet, following the user pressing the ‘start’ button, to the time when a selection was registered by Pure Data. The task durations of all scoring trials were then summed over all target locations (including when the target was not present) within each onset asynchrony/word duration block for each participant. Trials in which the participant responded incorrectly or which were added as dummy trials were excluded from this sum. This effectively removed the nuisance variable ‘target location’ from the analysis, leaving each participant one total task duration for each experimental block.

A positive skew at one onset asynchrony/word duration combination was observed. Since this violated the normality assumption required for parametric analysis, the square root of the aggregated task duration data was used. A Shapiro-Wilk test confirmed that the transformed data was not significantly different from normal ( $p > .05$ ). A two-way repeated measures analysis of variance (rm-ANOVA) was performed with onset asynchrony and word



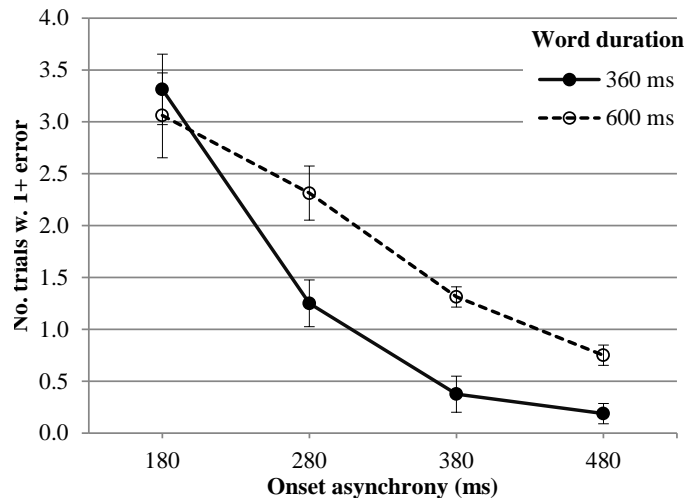
**Figure 5.7:** Marginal means of the square root transformed total task durations. The 360 and 600 ms word durations are represented by the solid line with filled markers and the dotted line with hollow markers respectively. (Error bars =  $\pm 1S.E.$ )

duration as independent variables.

Mauchly's test indicated that the sphericity assumption was violated for the onset asynchrony condition ( $\chi^2(5) = 13.5, p < .05$ ) and therefore it was decided to use the Greenhouse-Geisser correction ( $\epsilon = .60$ ). Sphericity was met for the word duration (2 levels) and the onset asynchrony  $\times$  word duration interaction ( $p > .05$ ). The results of the rm-ANOVA indicated significant effects for onset asynchrony ( $F(1.81, 27.1) = 8.79, p = .002, h_p^2 = .369$ ) and word duration ( $F(1, 15) = 25.3, p < .001, h_p^2 = .627$ ), while the interaction was found to be non-significant ( $F(3, 45) = 1.19, p = .323, h_p^2 = .074$ ) (see Figure 5.7). *Post-hoc* pairwise tests were performed for onset asynchrony using a Bonferroni correction, which indicated that the 280 ms onset asynchrony conditions led to significantly shorter total task durations than the 180 ms condition ( $p = .038$ ) and the 480 ms condition ( $p < .001$ ). The total task durations were also found to be significantly shorter in the 380 ms condition than the 480 ms condition ( $p < 0.001$ ). All other comparisons were found to be non-significant ( $p > .05$ ).

### Error rate

The error rates were taken as the number of target locations which required one or more repeats per block (including the not-in-list option) (see Figure 5.8). Statistical analysis was performed using generalised estimating equations (GEE) (Zeger & Liang, 1986). GEE analysis was chosen because the observed error rates violate the assumptions of normality required for traditional analysis of variance (ANOVA)-based methods. As the dependent



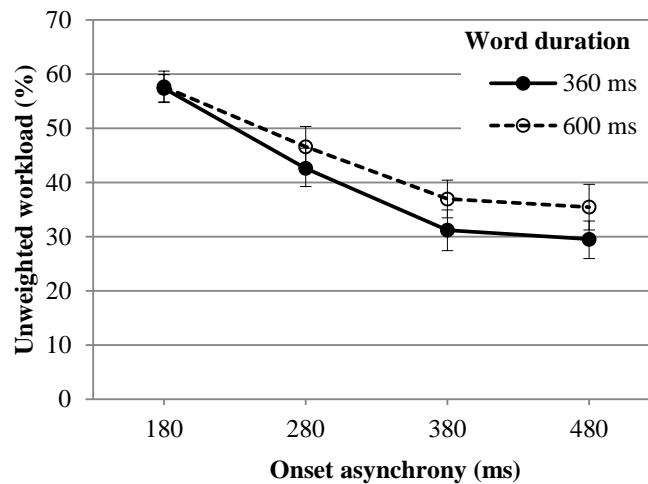
**Figure 5.8:** Marginal means (original scale) for the number of trials requiring one or more repeats during each block of 10 target locations. The 360 and 600 ms word durations are represented by the solid line with filled markers and the dotted line with hollow markers respectively. (Error bars =  $\pm 1S.E.$ )

variable was count data, the model was constructed using a Poisson distribution and a log-link function. The working correlation matrix was specified as auto-regressive (AR(1)) because error rates were likely to be more correlated with neighbouring onset asynchrony/word duration conditions. Convergence criteria were set as an absolute difference between iterations of less than  $10^{-6}$ .

The model fit values were 118 and 122 (to 3 s.f.) for the quasi likelihood under independence model criterion (QIC) and the corrected quasi likelihood under independence model criterion (QICC) respectively. Results of the model indicated that the effects of onset asynchrony ( $Wald \chi^2(3) = 113, p < .001$ ), word length ( $Wald \chi^2(1) = 26.7, p < .001$ ) and their interaction ( $Wald \chi^2(3) = 37.0, p < .001$ ) were significant. *Post hoc* Bonferroni-corrected pairwise comparison of the interaction indicated that for the 360 ms stimuli all onset asynchronies were significantly different from each other with the exception of the 380 and 480 ms conditions. For the 600 ms word duration stimuli no adjacent onset asynchronies were found to provide significant improvements, although each condition was found to be significantly different from all others. Word durations were significantly different for the same asynchrony only in the 380 and 480 ms asynchrony conditions.

## Workload

As one aim of the study was to investigate the effects of the display variation on overall workload, a detailed analysis of the subscales was not deemed necessary. Analysis, therefore,



**Figure 5.9:** Marginal means for the unweighted TLX scores for the 360 (solid line with filled marker) and 600 ms (dashed line with hollow marker) word durations. (Error bars =  $\pm 1S.E.$ )

focussed on the aggregated unweighted scores.

As in the pilot, the unweighted scores produced by the TLX software (Cao *et al.*, 2009) were used (see Figure 5.9). Shapiro-Wilk and Mauchly tests indicated that the normality and sphericity assumptions were met ( $p > .05$ ). Results from an rm-ANOVA (onset asynchrony  $\times$  word duration) indicated a significant main effect ( $F(3, 45) = 36.3, p < .001, h_p^2 = .708$ ) for onset asynchrony but no significant effect for word length ( $F(1, 15) = 3.43, p = .084, h_p^2 = .186$ ) or the interaction ( $F(3, 45) = .617, p = .608, h_p^2 = .040$ ). *Post-hoc* Bonferroni-corrected pairwise comparisons for onset asynchrony indicated that all of the treatments were significantly different from each other, with the exception of the 380 and 480 ms conditions.

### Post-experiment interview

General comments on the experiment indicated that the participants considered the conditions to represent a wide range of difficulties (e.g., “some of them were far easier than others” [P15]). A few participants also spoke about feeling as if they were improving through practice (e.g., “it was hard at first but then with practise I found it got easier [...] it felt like a more natural thing to do once you’d done it several times” [P11]). One participant commented on the interface’s pads not being as responsive as they could be (“I don’t think that’s the most responsive pad” [P13]) and that the voice sounded artificial (“I think because it’s a computer generated voice you can’t tell if it is saying heave in a diff... in a funny f... way” [P13]). Three of the participants commented that they found words in the right/final position

easier than those in the other locations (e.g., “I felt like I was performing less well on the the middle, front, hearing the the middle one less well. Um. Obviously the right-hand one is easier because it’s happening last so err. . . picking out the endings of word is is easier” [P6]). One participant mentioned the fatiguing effect of the experiment, saying “after doing it for a long time you would get fatigued cos kind of like there were times especially after the break where I felt pretty good and then towards the end of the sessions I’d be, especially when it was faster, I’d be wavering” [P12].

When participants were asked whether they had a preference for words being presented one at a time or with some overlap, reactions were mixed. While some participants favoured a serial display, others felt that some overlap was preferable. One participant suggested that the overlap could be something that was introduced when a user was more familiar with the interface (“if it was the first time I was using something, err. . . word at a time. If, however, it was a common user-journey that I was making repeatedly [...] where it was repeating something that I already would know of, then I would be more comfortable with overlapping” [P3]). Another favoured a display in which the words were distinct but had a short inter-stimuli interval (“one at a time but very close together” [P15]).

In terms of strategies, most of the participants reported attending specifically to the ends of the words in order to detect the target (e.g., “just listening to the end because they all started the same. So it would just be the last consonant that changes” [P0]). Some participants went into further detail and mentioned adopting different strategies depending on the presentation (e.g., “I think in each set I had to adapt it, so I. . . I couldn’t. . . I didn’t feel like I applied the same approach to every set or every test that I did.” [P3]). This reflects an issue with the format of the interview. As participants were only asked at the end of all of the trials, information about how strategies varied over the conditions is lost. It seems likely that strategies will have varied considerably over the different experimental conditions.

Interestingly, one participant reported that “not listening seems to work best?” and “Not trying to pick things out and kind of respond unconsciously rather than consciously seemed to help sometimes.” [P1]. A few of the participants mentioned revisiting their memory to complete the task (e.g., “with the overlapping ones [...] I replayed the thing afterwards in my head” [P12]). One participant spoke of reorienting their visual attention while listening to the words (“I think I noticed myself a few times kind of moving my eyes from there. . . from like the left hand side to the middle to the right hand side. Kind of trying to correspond with where the words were so that I can kind of figure that out as well” [P14])



The degree to which participants reported being able to identify other words in the lists varied considerably. While some simply indicated that they could, others indicated that it changed throughout the conditions or that they were often unable to (e.g., “when they were quite spaced out and spatially and temporally distinct, that was a bit easier. But yeah, when they were kind of overlapping and mashed together it was almost impossible.” [P1]).

One participant commented that the words were more different in some of the trials (“in the last trial that you’d mixed up the words? But on other ones they sounded like they were all si...similar but changing on just t...one or two of the letters” [P3]). It is unclear what the participant was referring to here as the stimuli all came from the same corpus. It is possible, however, that they were referring to the longer word condition, which they heard in the second experimental session.

When asked about how the longer words compared to the shorter ones many of the participants reported either not noticing a difference (e.g., “I couldn’t tell” [P15] and “I can’t answer that because I don’t remember it happening” [P13]) or not thinking it had an effect (e.g., “I don’t think it made much difference.” [P9]). Some participants expressed a preference towards shorter words. A couple of participants mentioned difficulties that could have been due to the timescale manipulation (“it made it more difficult in all cases um...they seemed unnaturally laboured and so even when they were distinct it...they were they were still harder to perceive as words because they were *very slowly spoken* [spoken in a slowed-down, drawn-out manner] which is weird” [P1] and “it sort of seemed like the words blurred and I had diff...more difficulty selecting” [P4]). While another referred to the presence of an overlap in the conditions making it harder with longer words (“words being longer made it more difficult um...because there was that little bit of an overlap I think between them, and it made it a bit more difficult to try and figure out when one word stopped and when another word began.” [P14]). One participant indicated a preference towards the longer words, stating simply “I think it made it easier” [P10]. Another suggested that it depended on the words “I think for shorter words when they were quicker and snappier I found that easier, but then for longer words when they like were more drawn out I found that easier” [P8]. A participant talked about there being “more phonemes to hook onto” [P3] in the longer words, which would not have been the case in the trials. One of the participants seemed quite puzzled by the question, as though they were trying to apply reasoning in their answer rather than basing it on their experience.

### 5.4.3 Discussion

The results presented in the previous section demonstrate significant effects on the task durations and accuracy from different onset asynchronies and word durations. Additionally, workload ratings were significantly affected by the onset asynchrony. This section interprets these statistical differences alongside the qualitative data, and findings of other studies. Based on these, the most appropriate display methods for use in spoken auditory menus are determined.

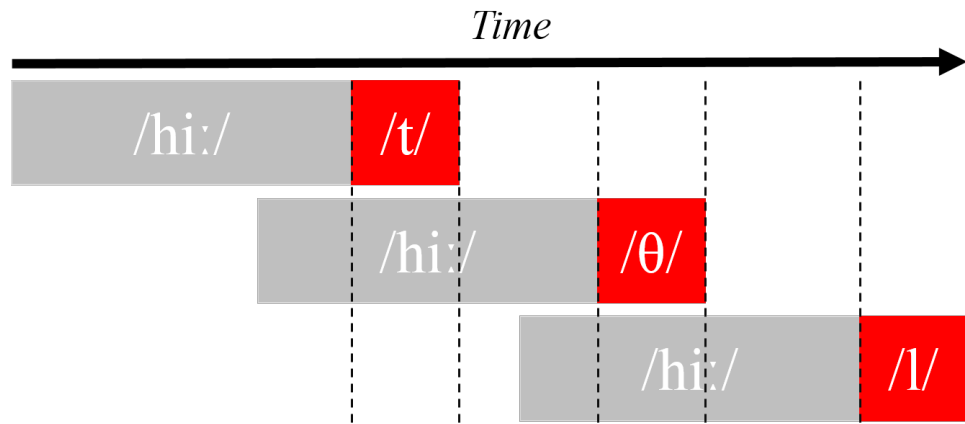
#### Task duration

Due to the implicit effect of shortened words on the time taken to present information, it is of little surprise that the word duration factor exhibited a large significant effect on task durations. The *post hoc* analysis of the effect of the onset asynchrony on the total task duration indicates an optimum asynchrony of around 280 ms, despite this representing considerably different durations of overlap between the two word duration conditions.

The lack of a significant interaction between the word duration and onset asynchrony conditions suggests that the degree of asynchrony, as opposed to the proportion of overlapping stimuli, was most important in determining the time taken on each task. For the onset asynchrony conditions above 280 ms it is possible that reaction speed advantage is present due to the words being more easily identifiable. As the task durations increase, however, any improvement in time taken to detect the target is less than the increase in presentation time when the asynchrony is at its maximum of 480 ms.

#### Error rates

The observed interaction in the error rates appears to be due to diverging error rates for the two word duration conditions as the asynchrony increased, with the difference becoming significant for the 380 and 480 ms asynchrony conditions. At these asynchronies the shorter stimuli are no longer overlapping, whereas the longer stimuli still overlap with the following word. This fact is particularly pertinent when it is recognised that the overlaps involve the endings of two words in each triplet, which in this task can be seen as the critical section for distinguishing between the maskers and the target word. This contrast in acoustic conditions was evidently more significant than between varying degrees of overlap in the smaller asynchrony conditions. It is, however, notable that the difference between the word



**Figure 5.10:** Visual representation showing that the critical phonetic information for the first two words (*heat* and *heath*) overlaps with other concurrent speech, while for the third word (*heal*) it does not.

durations in the 280 ms asynchrony conditions is considerable and it is speculated that an increased sample size might have led to significance. The influence of this overlap for the first two words is further emphasised by several of the participants identifying the left/first and central/second words in the display as being more difficult to resolve (see Figure 5.10).

The fall in error rate data over onset asynchrony appears generally in agreement with results from other studies on onset asynchrony (Lee & Humes, 2012; Ikei *et al.*, 2006). However, error rates appear to be higher than those found by Ikei *et al.* (2006) in equivalent conditions. Whilst it is possible that the use of intensity panning rather than binaural spatialisation may have contributed to the decreased accuracy, it seems unlikely that this difference alone would cause such a large discrepancy. It is also possible that modification of the word duration and pitch may have affected word intelligibility and inflated error values. From one of the comments made in the interview, it is clear that modified voice was not considered to sound natural. A couple of participants spoke of having more difficulty with longer words due to their timescale modification, but many participants did not notice there were two different lengths of stimuli or said it made no difference. Furthermore, it can be seen for both word durations that the accuracy approached 100% as the words became temporally distinct, implying that the processing of the words was not a major factor in itself. It is therefore likely that the difference between the error rates found here and those found by Ikei *et al.* (2006) is predominantly due to the choice of experimental stimuli within this study. Whilst stimuli in this trial were distinguishable through only the final vowel-consonant transition, the words used by Ikei *et al.* (2006) were more phonetically varied. This will have made their tasks considerably easier, as the increased phonetic variation will have provided the

participants with more cues by which to distinguish the target word from the maskers.

It is possible that the increase in error rate observed here is responsible for an apparent disparity between the trend in error rate found by Ikei *et al.* (2006) and the one found in this study. Whilst the results from (Ikei *et al.*, 2006) appear to show an optimum onset asynchrony of 300 ms (for three voices and no attenuation), the results of this experiment appear to show reducing error rates up to the 480 ms condition for the longer stimuli. It is thought that this inconsistency is a by-product of the inflated error rate present in this study, and therefore the optimum asynchrony suggested by Ikei *et al.* (2006) is the result of a floor effect on error rates. Whilst Ikei *et al.* (2006) indicates that greater accuracy could be achieved through the addition of ‘cross-ordering’ (presenting each word on the contralateral hemisphere to the preceding word) and applying an attenuation over the course of the word, neither of these methods were included in the design of the present study. Cross-ordering would not have been applicable due to the use of only three overlapping sources. It is feasible that through improving the audibility of word onsets, attenuation processing could have improved stream formation. In scenarios where the critical information is at the end of the word, however, the reduced SNR is hard to justify.

Research into backwards recognition masking (BRM) indicates that vowel recognition performance plateaus when vowel onsets are separated by 200-250 ms or greater (Massaro, 1974). The range of asynchronies in the present study suggests that BRM is unlikely to have been an influential factor for any asynchronies other than the 180 ms treatment. Due to the non-stationary nature of the speech signals used here, it could be that BRM impacted the stream formation and therefore made the location of the target more challenging to resolve. From the post-experiment interviews it is clear that the constant location of the critical information led many of the participants to attempt to attend to the word endings in order to detect the targets. If they only heard this element, participants will have had to use ordinal, spatial and, if the word ending is voiced, pitch information to derive which of the three locations the target had originated from.

## **Workload**

The results of analysing the workload scores indicates that onset asynchrony was the only factor that influenced the participants’ perception of task difficulty. In fact the workload scores appear to exhibit a divergent behaviour similar to error rate, though this difference was not large enough to be significant. Interestingly, this implies that the additional overlap

associated with the longer stimuli did not significantly contribute to participants' subjective workloads in the two largest asynchronies, despite significantly increasing their error rate.

### **Overlap or onset asynchrony**

It would appear that onset asynchrony describes observed trends for task duration and workload better than the amount of overlap. The error rate, however, displays a more complex interaction between the onset asynchrony and word length. The divergence between word durations with increasing asynchrony implies that both asynchrony and overlap influence performance. It is acknowledged that the difference between word durations was comparatively small due to the nature of the stimuli chosen and based on this study it is not possible to come to any conclusion regarding situations in which the amount of overlap is considerably larger.

### **Asynchrony in menu display**

Considering the effects of asynchrony on navigational speed, accuracy and subjective workload, it would appear that, of the treatments measured, the onset asynchrony of 380 ms provides the best compromise across all performance measures. In terms of task duration alone, the lack of a significant difference between the 280 and 380 ms asynchrony conditions implies that an optimum exists between the two measured treatments. If one considers the additional time that would be incurred due to the higher error rates associated with the 280 ms condition, it seems likely that in practice this optimum is closer to the 380 ms condition. This conclusion is supported further through the workload scores, which show a significant reduction in workload from 280 to 380 ms onset asynchrony, suggesting that users felt that this condition made the interface significantly easier to use. The lack of overlap for this asynchrony condition for the shorter words, and its effect on error rate and navigational speed, is particularly pertinent, as it suggests that a more efficient solution would be to temporally compress the stimuli and present them with a short inter-stimuli interval. Interestingly, Werner *et al.* (2015) found that temporally compressed serial stimuli led to slower menu navigations than were found when normal speed speech was used and were associated with worse accuracy rates than normal speed serial speech and concurrent presentations. As noted by Werner *et al.* (2015, p. 1098), this may have been due to the lack of experience the participants had with temporally compressed speech.

Use of a non-overlapping display raises a question over whether the grouping of stimuli

into triplets is advantageous. Grouping would seem likely to increase speed, as the number of physical interactions with the interface are reduced. Previous work comparing grouped and individual presentations of temporally distinct spoken items, however, indicates that participants are able to navigate to target locations faster when words are presented one at a time (Sodnik *et al.*, 2011). The grouped display in (Sodnik *et al.*, 2011) imposed 200 ms inter-stimuli delays, whereas the present study, when using the shorter stimuli and the 380 and 480 ms asynchronies, creates inter-stimuli delays of 20 and 120 ms, respectively. This suggests that faster navigation may have been possible by reducing the size of the inter-stimuli delay with minimal impact on workload and error rate. Further advantages of grouping displays, rather than requiring users to navigate through individual items, are that following the results of (Ryan, 1969) it seems that it will be easier for users to remember the items, which may provide benefits for repeated use. Furthermore, one-by-one navigation requires users to make a simple decision for every single item in the display rather than fewer more complex decisions, as would be the case in a grouped display. This could be advantageous, as making few complex decisions is found to be more efficient than making many simple decisions (referred to as decision complexity advantage) (discussed in Wickens *et al.* (2016)). To inform the future design of spoken auditory displays, further investigation is recommended to ascertain the effect on performance of grouped displays with lower inter-stimuli delays.

It is interesting that there was a split in the participants' opinions over whether they would prefer to use serial or overlapping speech displays on a regular basis. This appears to indicate that some concurrency is not viewed as being entirely negative, even though the performance metrics considered did not find it to be optimal. It is notable that the questions were posed at the end of the study and trial conditions were not explicitly defined to the participants. This means that robust inferences from this result cannot be drawn. To gain better insight into this, it would be necessary to measure preferences throughout the experiment.

The methodology presented here primes the user with a visual representation of the target word and it simulates a user with a very clear idea of the item which they are looking for. Where the target is initially known, a search-based navigation is likely to prove more efficient. The present methodology is a selective attention task in which the user need only listen out for one word within the list and can ignore all others. This is distinct from what is required in a browsing task where a user would be expected to have to listen to a set of possible selections before making a choice. The methodology in the present study was adopted to reduce response variation due to possible target identity confusion and therefore represents

the ideal scenario in terms of target knowledge. In the post-experiment interview, participants indicated having some awareness of other words in the lists at some points in the experiment. Without quantitative testing of performance at recalling the full lists, however, it is not clear how individual treatments affected this ability.

It is argued that the results of this experiment are transferable to hierarchical menu structures, as these navigations comprise many one-dimensional navigations like those tested here. Any additional testing with hierarchical menus would, however, require a very different experimental methodology. Restricting the stimuli based on their phonetics would be problematic, as it would be practically impossible to use terms with meaningful hierarchy. While it would be possible to simulate hierarchical navigation by giving a participant a list of target words (e.g., “Find *bin*, then *sub* etc.”), this introduces the potential for participants to become confused about which target they are looking for at a specific moment.

It is worth noting that the stimuli used within this study were quite short, which may have restricted the degree of stream formation that could occur, causing critical information to be missed, or its location/order/pitch to be unresolved. With longer, less informationally dense content, as in (Guerreiro & Gonçalves, 2014, 2016), users may have been able to orientate their attention more effectively towards a desired stream of speech. It is unclear whether this would offer a significant advantage in terms of both time saved and accuracy.

## 5.5 Summary

The problem of providing users with non-visual menus capable of facilitating fast and accurate navigation is a considerable design challenge using an auditory display. Due to the limitations of non-speech methods regarding the representation of dynamic, novel content, it would appear that speech-based methods are most appropriate. The experimental work described in this chapter explores the feasibility of using asynchronous, overlapping speech for menu representation and seeks to determine what effect this has on the speed of navigation, accuracy and workload.

The design of a display for menus using asynchronous spoken items is discussed which employed a group of three talkers with different voice pitches and uttering single words with varying onset asynchronies. The talkers were rendered in spatially distinct directions using intensity panning. Consideration is given to both sliding-window and segmented navigational models. Segmented displays that allow a user to select any of the concurrent items are

identified as having the potential for facilitating the lowest navigation times. For this reason, the menu design proposed here follows this model.

An experiment was undertaken in which participants attempted to find a target word within a list of words. Task duration, accuracy and subjective workload were assessed for different onset asynchronies and word durations. The results of this experiment indicate that performance metrics and workload ratings at the lowest onset asynchronies (180 ms) were improved by increasing the amount of onset asynchrony. Regardless of the word duration, the lowest navigation times were observed to be between the 280 and 380 ms onset asynchronies. Taking into account the lower error rates associated with the 380 ms asynchrony, it is suggested that performance in real-world navigation tasks would be best at this setting. Furthermore, workload ratings were lowest for the 380 ms condition and were not significantly improved by additional asynchrony. An interesting finding is the lack of a significant effect in workload ratings for the word durations, despite the conditions having a considerable impact on the amount of temporal overlap between stimuli. This implies that the onset asynchrony was a more significant factor in terms of the user experience than the amount that stimuli overlapped.

Interestingly, at the optimum onset asynchrony of 380 ms the error rates were significantly lower for the shorter stimuli, which no longer temporally overlapped with each other. This is taken to suggest that, although speed, accuracy and workload advantages can be observed when words overlap, a better approach may be to present shorter or temporally compressed words grouped into triplets with a short inter-stimuli interval.

While this experiment points towards a serial presentation as being best within this use case, the duration of the stimuli and high density of critical information may have made this task particularly difficult. In this study, the nature of the task also demanded that participants were unable to predict which stream they should attempt to attend to. The cost of orienting attention to the incorrect source could, therefore, have been an important factor when words were overlapping. Furthermore, success within this context is measured by the time taken to accurately complete a task. It is clear that, while this is an important metric for menu navigation, there are other use cases around the connected television experience where this is not the case. It follows that some concurrency may prove beneficial in other television use cases that will be considered in the following chapters.



## Chapter 6

# Orchestrated synchronous companion experiences: Non-visual design considerations

### 6.1 Introduction

Companion experiences for television have been introduced as one of the emerging user experiences for connected television. Much of the work in this area has concentrated only on using the screens of secondary devices to present additional content. Most of these systems, however, also have audio playback functionality and some have haptic capabilities too. It is the intention of this chapter to explore how the use of additional auditory channels alongside television programmes may provide new user-experiences for television and how inter/intra-modal concurrency may be used.

According to the taxonomy laid out by (Hoare & Hinde, 2016), previously introduced in Chapter 2, this work will focus on orchestrated synchronous companion experiences which are fixed and scheduled. This chapter refers to elements of an orchestrated synchronous companion experience according to the terms in Table 6.1. This type of experience is particularly interesting as the user divides their attention between an on-going programme and the additional activity, and the orchestrator exerts maximal control over the experience. Due to the degree of control taken by the orchestrator, the design of these experiences requires a great degree of care. The factors governing the success of these experiences, however, are not well understood. Furthermore, while visual experiences have been developed, there have

**Table 6.1:** Tables of terms used to describe different components of a scheduled orchestrated companion experience

Acronym	Description
MPC	main programme content— <i>all content that is intrinsic to the programme. In the case of television, it is considered to be bimodal (i.e., the audio and visual components of a traditional television programme including accessibility features).</i>
SPC	secondary programme content— <i>content that is designed to be displayed in an orchestrated synchronous companion experience and does not form an intrinsic part of main programme. Considered here as unimodal. (e.g., text or speech describing additional information about the location of a scene)</i>
VSPC	visual secondary programme content
ASPC	auditory secondary programme content

yet to be any auditory equivalents.

This chapter focuses on use cases in which the main programme content (MPC) is bimodal comprising both visual and auditory components, while the secondary programme content (SPC) considered in this section is unimodal either visual or auditory. It should be recognised that visual SPC experiences comprise scenarios in which the additional media is experienced on a different screen and scenarios in which additional content is represented on the same screen as the MPC. While many of the principles discussed in this chapter are likely to impact both of these scenarios, it is notable that the focus of this chapter is on visual secondary programme content (VSPC) experiences on secondary devices. Displaying VSPC on the same screen as the MPC requires compromises to be made to the visual MPC whereby elements are either occluded by SPC overlays or the resolution of the MPC is modified so as to make space for SPC. This, in turn, has important ramifications for shared television experiences, as one user's desire to access additional content impacts other viewers' experiences of the MPC. Accessing content on an additional screen provides a simple alternative which bypasses these issues. Furthermore, with the popularity of personal connected devices, it seems that this option is likely to become common in future visual experiences.

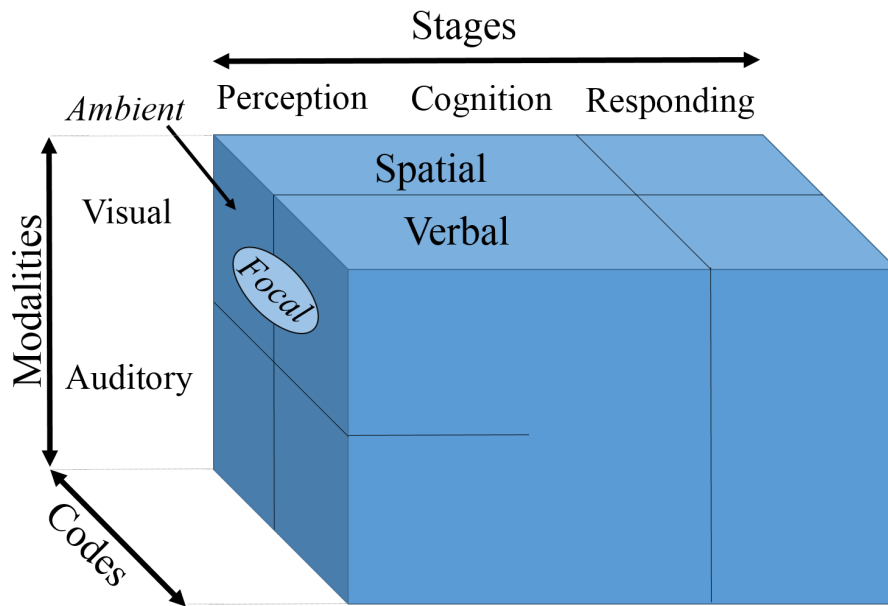
Within this chapter, first, the findings of work concerning the design of visual synchronous companion experiences are considered. Then, important factors to be considered for the design of an auditory display are discussed, with reference to related work from researchers

considering audio description (AD).

## 6.2 Attention with visual secondary programme content

The visual presentation of SPC raises interesting questions regarding how users direct their attention. The user must choose whether to look at the secondary device or the main screen. As part of this decision, the user must sacrifice the information on the other, unattended screen. Researchers have sought to understand how users divide their attention in these scenarios, either using eye-tracking (Holmes *et al.*, 2012; Brown *et al.*, 2014) or video analysis of gaze direction (Dowell *et al.*, 2015). The manner in which users choose to distribute their attention across screens may be affected by their level of interest in the content on the main screen. Brown *et al.* (2014) noted in their study using a nature magazine show (“Autumnwatch”) that participants were more likely to view the second screen when the programme cut to the studio rather than during segments shot on location. In this context, users were apparently less visually engaged by the studio footage than by the prospect of the secondary content. Dowell *et al.* (2015) suggest, however, that the content of the auditory channel may also be a factor in when users choose to orient attention to the secondary device. In traditional television or film production, the audience’s attention is directed by the programme maker using auditory and visual cues (i.e., focus, lighting, framing etc.) (Carroll, 2003). Giving the audience a decision between two concurrent streams breaks this tradition and introduces the possibility of missing information. This decision itself can introduce difficulties with users not knowing where to look, as was reported by Dowell *et al.* (2015). Furthermore, the issue of how to notify the user of the presence of new VSPC is introduced. Neate *et al.* (2015) compared the use of notifications on the main screen (still or shaking) and non-speech auditory notifications (earcon or auditory icon) with content appearing on the secondary device (still or shaking). Earcons were found to elicit the shortest reaction times, significantly lower than the visual notification on the main screen and were considered more effective at getting attention to content appearing with no notification. Participants’ preference ratings, however, indicated that auditory icons and the TV-based alerts were preferred to the condition with no alert.

While the cost of diverting visual attention on the visual information from the MPC is apparent, the effect of VSPC on the information from the soundtrack is less clear. The multiple resource theory of attention suggests that the degree of interference that occurs



**Figure 6.1:** Multiple resource model of attention (adapted from Wickens (2002, p. 163))

between concurrent tasks is governed by the extent to which they draw from the same pools of resources (Wickens, 1980, 1984). The model of multiple resources has four dimensions, which comprise: modality, stage of processing (perception, cognition or response), coding (spatial or verbal) and visual channel (focal or ambient) (Wickens, 2002) (See Figure 6.1). It follows that the modal separation between the visual SPC and the auditory MPC soundtrack does not implicitly rule out interference between the streams. In fact, it seems likely that the amount of interference that occurs is dependent on the nature of the information present in the MPC and SPC at any time. VSPC is likely to comprise text and/or images. Whilst the MPC soundtrack is likely to comprise speech, music and sound effects (i.e., sounds of on-screen action and ambient sounds in the scene) (Butler, 2007). Butler (2007) describes four purposes of television audio: attracting and maintaining the audience's attention, aiding understanding of the narrative and meaning, creating smoother transitions when cutting between scenes, and providing the illusion of continuity across shots within a scene that have been shot non-sequentially. The type of information conveyed by sound, its importance and the type of mental processing it demands are therefore dependent on the type of sound and its usage. For example, an on-screen conversation between two main characters may convey important plot information and require the audience to process the speech and comprehend it. Conversely, off-screen speech of minor characters in a scene may be present to convey context (e.g., background chatter in a busy cafe) and is required only for the audience to perceive its presence.

The presentation of images alongside music, speech and effects is a device commonly exploited within the televisual experience. From the multiple resource model of attention (Wickens, 1980, 1984, 2002), the benefits of this bimodal presentation are argued to arise from the reduced cognitive demands compared to speech and the streams not drawing on common resources. It is therefore assumed, when displaying a slide show of photos on a secondary device, that the two streams of information do not impede each other. Care is necessary here, as this is not to say that the two streams do not have an effect on each other. The editorial use of sound in television to support, contradict, or emphasise elements of the footage (Butler, 2007) clearly displays the potential of sound to impact a user's interpretation of an image.

Of the possible scenarios posed by VSPC experiences, the presentation of additional text alongside MPC dialogue or narration seems most likely to lead to disruption. Both speech and text are verbal codes and require cognition for processing. Multiple resource theory would therefore suggest that a significant amount of interaction is to be expected. In the study of interactions between visual processing of textual information and verbal stimuli, many authors have made use of list recall measures (e.g., Salamé & Baddeley, 1987, 1989; Jones *et al.*, 1992; LeCompte, 1994). This, however, is not necessarily representative of the process that would be utilised during normal reading (Baddeley, 1997). Experiments in which participants have been asked to process different speech and text stimuli simultaneously appear to show that participants are unable to attend to both streams (Mowbray, 1953, 1954). It seems likely, therefore, that when users are reading textual material in SPC they are choosing to process the textual materials at the cost of the concurrently presented speech. Even in selective attention tasks where participants have been asked to read text in the presence of irrelevant speech, researchers have found detrimental effects on comprehension (Martin *et al.*, 1988; Oswald *et al.*, 2000; Sörqvist *et al.*, 2010) or, when time was not limited, increased reading times (Cauchard *et al.*, 2012). Furthermore, work specifically focussing on the impact on reading of background television content containing a lot of spoken content has found detrimental effects on reading comprehension (Armstrong *et al.*, 1991; Armstrong & Chung, 2000; Ylias & Heaven, 2003). It would therefore seem that reading textual SPC during the presentation of MPC containing speech is likely to have a detrimental impact on the comprehension of the SPC compared to an asynchronous experience.

Though users are likely to be employing selective attention in concurrent speech and reading scenarios, some information from the ignored speech stream may still be processed. Findings that the detrimental effect of speech on reading comprehension is dependent on the speech

being meaningful implies some semantic processing is occurring (Martin *et al.*, 1988). Brown *et al.* (2014) report that users returned their gaze to the main screen after an exclamation in the MPC soundtrack. This could be seen as a sign that users were able to process speech from the MPC. It seems likely, however, that a salient auditory event such as a vocal exclamation would draw attention regardless of whether the rest of the speech was being processed. Furthermore, while textual content was present in the SPC, it was generally quite short and was presented alongside images. Demands on users' linguistic resources may therefore have been sufficiently small as to have had minimal effect.

The impact of music playing during reading is somewhat more complex. Firstly, in terms of dual task performance it is difficult to assess what informational content a listener in the real world would extract from music. There is, however, some evidence of added music altering emotional evaluation of the text (Cassidy & MacDonald, 2007). Considering the multiple resource model of attention (Wickens, 1980, 1984), one would not expect instrumental music and reading to interact to the same degree as speech due to the non-verbal nature of instrumental music. Studies on the impact of background music on reading have, however, still found detrimental effects from both lyrical (Martin *et al.*, 1988; Furnham & Strbac, 2002; Avila *et al.*, 2012; Perham & Currie, 2014) and instrumental music (Avila *et al.*, 2012) when compared with quiet conditions. Nevertheless, in comparison to speech, it seems likely that instrumental music is less disruptive to reading (Martin *et al.*, 1988; Cauchard *et al.*, 2012). This said, music is hugely variable in both its spectro-temporal characteristics and its affective qualities, which makes generalised comments on the effect of music problematic. For example, there is some indication that music which is more aggressive (Cassidy & MacDonald, 2007), or has a higher tempo and greater intensity (Thompson *et al.*, 2012), is more detrimental to reading.

Considering the effects and ambient content in a MPC soundtrack, there exists considerably less applicable research. The atmospheric components of the soundtrack face similar difficulties to music in that it is hard to characterise the information they convey or to generalise their acoustic characteristics. Research looking at the impact of environmental noise on reading has revealed some detrimental effects on reading which are similar to those caused by music (Furnham & Strbac, 2002; Cassidy & MacDonald, 2007). It is noteworthy, however, that these studies included some speech within the environmental noise (mumbling in the case of Furnham & Strbac (2002)). It is difficult to determine how applicable this would be to the atmospheric sounds in television. Also, as most reading occurs in the presence of

some ambient sound, it seems that the presence of some environmental noise is not overly disruptive to reading.

To summarise, it appears that with the presentation of textual SPC the user must decide whether to miss the MPC speech, or the information in the SPC. Furthermore, should they choose to devote attentional resources to the SPC, reading the content will be more difficult than if it was read in isolation.

### 6.3 Auditory presentation of secondary programme content

While there has been some investigation of the use of auditory notification for SPC (Neate *et al.*, 2015), there has yet to be any exploration of conveying SPC in the auditory mode. This thesis focuses on the creation of an auditory equivalent to additional textual descriptions in VSPC. Considering only textual passages of VSPC may be seen as an over-simplification of the type of information conveyed in VSPC experiences such as the “Autumwatch” companion, which also comprised images, animations, and diagrams (Jones, 2011). The scope of our work is limited in this way because the development of equivalent sonic representations of images, animations and diagrams is a complex problem in its own right. By focusing on the representation of textual passages, we aim to identify key design factors, which may be extended in later works to encompass these additional presentation modes. Through the exploration of this use case, our work further investigates the potential use of concurrent audio streams within consumer user experiences for television.

The previous chapter suggests that speed advantages offered by concurrent speech presentations for menus are unlikely to be better than using shorter/shortened words with no overlap. There are a number of key differences between menu representation and synchronous companion experiences, however, which suggest that concurrency may be more appropriate in this use case. Firstly, stimuli within orchestrated synchronous companion experiences will be considerably longer than those considered in menu navigation. This will give the user more time for stream formation and should, therefore, lead to improved selective attention performance.

Another important difference is the density of the information within concurrent streams. In the menu use case, the density of critical information means that any information missed from a stream could have a large impact on the success of the task. Within the context of SPC, however, the streams will contain considerably more redundant information, allowing

for information to be missed without impeding the user's understanding or enjoyment of the display. Furthermore, to detect the target within the menu display a user has to monitor information from all streams to detect a specified word. Within the context of synchronous companion experiences there is no such requirement for the user to split their attention in this way. This reduces the amount of information the user is required to process and can be expected to lead to a lower perceived workload. There is also a difference in what the serial equivalent would comprise. In terms of menu representation, this is simply a case of presenting the items in order so that they do not overlap. With a television programme, however, one serial approach may be to pause the programme during the SPC, while another may be to offer asynchronous experiences before or after the programme. Although these alternative approaches would allow access to the same information as a synchronous companion experience, they sacrifice either the timeliness of the information or the pace of the main programme. Clearly, there are considerable experiential differences between these presentations and visual synchronous companion experiences.

## 6.4 Design

The addition of auditory content to a completed television programme is a complex problem which requires careful design consideration. A television soundtrack typically contains combinations of distinct types of sound (e.g., atmospheres, Foley, music and speech), which pose different issues regarding distraction and masking. Furthermore, when the soundtrack contains important contextual and narrative information for the programme, it is important that the user's experience of it is not impaired. The challenges of adding extra auditory information to a television programme have been considered within the context of the original AUDETEL (AUdio DEscribed TELEvision) project, which developed the audio description (AD) system used in the UK (Lodge & Slater, 1992). AD consists of extra spoken descriptions of on-screen elements or actions during gaps in the programme dialogue. The AUDETEL project sought to improve television experience for people with visual impairments or who were blind (Lodge & Slater, 1992; Pettitt *et al.*, 1996). Since then, it has also been found to be beneficial for other groups (Incorporated Television Company (ITC), 2000; Fellowes, 2012; Walczak, 2016; K. Krejtz *et al.*, 2012; I. Krejtz *et al.*, 2012).

K. Krejtz *et al.* (2012) and I. Krejtz *et al.* (2012) investigated the potential use of AD to enhance the educational value of videos. They did this by adding AD to an animation and then measuring the ability of children to recognise scenes from the video, and to recall



information in the programme. They also captured eye-tracking data from the children while they watched the scenes. While the AD described what was happening in the scene, it was also used to communicate information that was not present in the original programme. Results concluded that the participants who had watched the videos with the AD gave better answers to questions about the contents of the clip (e.g., using correct terms from the AD) (I. Krejtz *et al.*, 2012). Analysis of the eye-tracking data revealed that the presentation of AD led to differences in eye-gaze behaviour. This indicated that the AD encouraged participants to focus more on relevant parts of some scenes (I. Krejtz *et al.*, 2012).

The description of information included in the AD by K. Krejtz *et al.* (2012) and I. Krejtz *et al.* (2012) sits between definitions of AD and auditory secondary programme content (ASPC), as the added speech both reinforced visual information from the MPC and added some new informational content. This and the specificity of the use case make it difficult to inform work on the design of ASPC in normal television watching scenarios. Whilst I. Krejtz *et al.* (2012) performed some comparison between presenting AD in the scene prior to the action, and presenting it alongside the scene, the work followed the display traditions of AD and alternative display methods were not considered.

The work into the provision of AD has explored interesting elements of user experience, which are likely to be pertinent to the presentation of ASPC. There are, however, considerable differences between the nature of information conveyed in these two applications, and with their relationships to MPC. A number of design decisions must be made in the creation of an ASPC display system (see Table 6.2). This section discusses and considers each of these factors.

#### 6.4.1 Content representation

The first factor that must be considered in the design of a display for ASPC is how the additional information should be represented. Work on AD highlights the issue of finding appropriate dialogue gaps in MPC to accommodate the required spoken description within television programmes (e.g., Chapdelaine, 2010; Encelle *et al.*, 2011, 2013). With the presentation of ASPC, this is likely to be even more of an issue, as SPC elements are likely to comprise longer textual descriptions.

Encelle *et al.* (2011) propose the use of earcons alongside speech for AD to allow more information to be conveyed in these gaps. Six earcons were used to represent the scene locations within the videos. Speech was used to describe all other elements that would

**Table 6.2:** A table outlining the design factors and questions that will be considered for the design of a scheduled orchestrated experience using audio.

Factor	Design questions
Content representation	Should ASPC be speech or non-speech?
Concurrency	Should ASPC occur at the same time as the MPC soundtrack or only elements of it?
Source characteristics	What type of talker should be used and how should it be recorded?
Interaction design	What should users control? How much interaction should they have? How can a display facilitate this?
Notification design	How should users be made aware of available SPC? What information should be conveyed and how?
Spatial configuration	Where should ASPC be presented from relative to the listener. Should MPC spatial configuration be altered?
Shared and individual user experiences	How can a display cater for individual and shared television watching with ASPC?

normally form part of the AD and introduce the earcons. User testing showed that the combination of earcons and speech could be useful, though some concerns were raised regarding the number of earcons that could be learnt and the effects of the additional audio on the ‘rhythm’ of the videos. While the approach by Encelle *et al.* (2011) demonstrates some benefit from the use of non-speech alongside television programme soundtracks for conveying information, the information being transferred was relatively simple and likely to be repeated several times within the show. With ASPC this is unlikely to be the case. Individual passages of SPC may have little that would be repeated and so little would be gained from this form of abbreviation.

Completely non-speech representations of the SPC may allow sufficient temporal compression to utilise these dialogue gaps. Restrictions associated with non-speech representations include the need for the representations to be learnt, resulting in a limit to the number of cues that can be provided and an inability to convey novel information. These constraints make non-speech cues impractical in this context. We conclude that speech is the only auditory code that would be practically capable of representing the required information and hence will form the basis of the proposed system for presenting the ASPC.

### 6.4.2 Concurrency of presentation

The addition of ASPC to MPC raises interesting questions about the amount of concurrency that should be present in the experience. The notion of a scheduled experience implies that the timing of the SPC relative to the MPC is of some importance. But how can auditory information be delivered in a timely manner alongside a complete television programme soundtrack?

A similar problem is faced in the presentation of AD, where much of the contextual information is related to specific scenes, or activities. With AD, describers in the UK are instructed to avoid overlapping AD with dialogue or narration in the MPC, and to reduce the sound level of the programme audio during delivery of the AD (Incorporated Television Company (ITC), 2000). The restriction of AD presentation to gaps in dialogue can be problematic. Descriptions are likely to be delivered some time before or after the corresponding visual information is displayed. Furthermore, the amount of description that may be given is limited by the duration of the gaps that occur in the dialogue. If the potential addition of AD were considered throughout the production process, it is conceivable that shows could be structured to allow the presentation of adequate, timely AD. In practice, however, this would likely disrupt the pace of delivery for the audience experiencing the show without added AD.

Some authors have sought to solve the timing limitations imposed on AD by pausing the main programme when the duration required by the description exceeds the available dialogue gaps (Chapdelaine & Gagnon, 2009; Encelle *et al.*, 2013). Chapdelaine & Gagnon (2009) propose a two tier system in which users can opt either for standard descriptions (fitted to the dialogue gaps) or extended descriptions that pause the video to accommodate their additional duration. Though predominantly positive user opinions on the overall system were recorded, it does not appear that any direct comparison was performed between the two levels of AD. A further study by Chapdelaine investigated the information which blind participants requested when watching videos without AD and suggested that accessible systems should provide a function that allows users to request confirmation of information in the MPC (e.g., characters' facial expressions or the cause of specific sounds) (Chapdelaine, 2010). An accessible DVD-player was created which featured the two-tier approach to AD and provided a 'recall assistance' function which allowed users to confirm understanding of elements within the scene (Chapdelaine, 2012).

Later work from Encelle *et al.* (2013) investigated the degree to which pausing videos to

facilitate additional descriptions caused users to feel discomfort. In the study, a series of audio-described videos were presented to blind participants, who were asked to press a button when they felt discomfort due to a pause. Participants reported a high number of discomforts on their first exposure to a video with the enhancements in place. The number dropped significantly, however, for the second video, suggesting that over time a user becomes used to the pauses. Interestingly, the length of the pauses was not found to significantly affect the number of discomforts reported. Overall, most participants' feedback on the use of these pauses was positive, though some participants suggested more fade-in/out on either side of the pause would have been an improvement.

The idea of Chapdelaine (2010) and Encelle *et al.* (2013) of inserting pauses to the MPC to facilitate the addition of spoken descriptions seems feasible in the context of the display of ASPC. Pausing live video streams is trivial within the context of IP delivery and even many STBs provide this functionality (e.g., Freeview, n.d.; YouView, n.d.). On the other hand, pausing exhibits some considerable limitations. Inserting pauses into a live stream means that the viewing is no longer truly live. This may seem acceptable when considering the audience as isolated groups, having individual experiences. This view is outdated, however, as the ubiquity of social media now connects users who can “enjoy the communal experience of group viewing without being physically together” (Wohn & Na, 2011). The delayed viewer finds out the identity of the killer, the final score, or who got eliminated after the rest of the audience after his/her fellow viewers. To be a delayed viewer in these scenarios is to be playing catch-up, understanding the context after the statement and unable to contribute to the conversation. It is notable, therefore, that social media “could well re-entrench synchronicity in television viewings, and make viewers less likely to use time-shifting technologies” (Harrington *et al.*, 2013, p. 407).

Through inserting pauses, the programme duration is extended. This idea of programmes without a fixed duration is, in itself, an interesting prospect being explored by researchers from BBC Research & Development with radio content (Armstrong *et al.*, 2014). While the work demonstrates the feasibility of this concept, the user experience implications have yet to be fully explored. Additionally, the SPC model would introduce further complications because duration would be liable to change throughout the programme, as a user chooses whether or not to access individual elements of ASPC. It is speculated that even with the application of variable-length programming, some users wishing to access elements of ASPC will not always want to extend the programme beyond its predefined duration. Furthermore,

the behavioural observations from usage of VSPC during commercial breaks (Holmes *et al.*, 2012) or when cutting to studio shots (Brown *et al.*, 2014) suggests that users may be using SPC at times when they are less interested in the MPC. Where this is the case, it would seem inappropriate to pause the MPC.

The discussion up to this point has considered a model in which each primary device is viewed by only one user at a time, or by a group of users all of whom want the same experience. This is, however, a clear over-simplification. It is likely that different users will have different interests and will therefore wish to access different elements of the ASPC. If not all parties engage with an element of the ASPC, then users with no interest in it will have their experience interrupted by a pause in the MPC, causing them disruption and frustration. Considering also the visual aspect of this experience, it seems likely that the stop-start nature of the experience would be particularly jarring.

The points raised above are not intended to imply that pausing MPC during the delivery of ASPC would not work in some cases, but serve to point out that a pause-based approach is unlikely to be appropriate in a number of scenarios. This project therefore considers only MPC with a fixed timeline, as is the case for orchestrated synchronous visual companion experiences.

The approach taken by AD is to minimise the concurrency between the added description and the programme audio, particularly for speech. This can be seen as sensible within the use case of AD for several reasons. Firstly, if AD were presented concurrently with programme dialogue the user would have to make a decision between selectively attending to the dialogue, selectively attending to the AD, or attempting to divide their attention between the two. If selective attention is used, the user chooses between information from the programme soundtrack and the AD. Either way, the user is engaged with the programme and is trying to follow the programme's narrative. Asking a user if they would rather know the context of the scene (e.g., where it is or who is there) or listen to the on-going dialogue, the likely answer is that they would want to know both. Attempting to divide attention, however, has been associated with a large reduction in the ability to correctly report information from the streams (Best *et al.*, 2006), which may mean that little information from either source is understood. If there are sufficiently large temporal gaps in which the AD may be delivered then this interleaved approach is sensible.

Secondly, it eliminates the occurrence of speech-on-speech masking. This is particularly pertinent when the demographics of people who are blind or visually impaired are considered.

The original work on the development of AD in the UK highlights the likelihood of people who are blind or have visual impairments also having some hearing impairment (Lodge & Slater, 1992). This observation provides the motivation for the ‘ducking’ of other programme audio when descriptions are added in the guidelines (Incorporated Television Company (ITC), 2000). The demographic of people with visual impairments or who are blind is skewed towards older people (World Health Organization, 2014). This point is important because research suggests that ageing is associated with a degradation in selective attention performance, which makes it harder to ignore unwanted signals (McDowd & Fillion, 1992; Tun *et al.*, 2002). Hence, masking incurred by concurrent streams of AD and dialogue would be excessively problematic for the target demographic.

For the use case of SPC these factors are less problematic. As previously stated, AD forms an intrinsic part of the MPC for its users, whilst SPC does not. Though it would be nonsensical for auditory description to obfuscate dialogue from the MPC, this is not necessarily the case for SPC due to its divergence from the MPC. This is particularly apparent when users are not attending to both streams concurrently in synchronous visual companion experiences (see discussion in Section 6.2). It is also true for someone without a hearing impairment, who is likely to suffer less masking by the AD than someone with an impairment. This said, avoidance of overlapping spoken ASPC with MPC dialogue may still be preferable due to the informational masking that may occur between concurrent speech streams.

Considering the difficulty in fitting adequate AD into naturally occurring dialogue gaps, it is unlikely that limiting ASPC to dialogue gaps would be achievable. Though pausing the MPC will not be considered further, it introduces the interesting prospect of modifying the MPC to facilitate the addition of the ASPC. This concept will be explored in more depth for models in which the MPC duration remains unaltered. The closest equivalent to pausing the MPC within this context would be to mute the MPC audio. This would remove any auditory masking between the MPC and ASPC streams and would therefore provide the best scenario for understanding the SPC from an acoustic perspective. From an experiential perspective, however, this is not necessarily optimal. Muting leaves the user completely unaware of the audio content of the MPC audio during this period, which may mean that they miss cues in the soundtrack regarding something they find interesting. It may also distance the user from the show’s rhythm and atmosphere.

Considering an OBB approach, as proposed by Armstrong *et al.* (2014), where the MPC is delivered as a collection of objects rather than as a single mix-down, more advanced MPC

modifications are possible. Having this extra level of control allows the system to selectively mute particular elements of the soundtrack during SPC presentation, which may otherwise have been disruptive. In this way, it is possible to blend the remaining elements of the MPC soundtrack to better integrate the SPC within the programme.

Speech in the MPC will be most disruptive to the understanding of the SPC due to the informational masking that it introduces. One possible approach, therefore, is to remove the speech from MPC whilst leaving the other elements of the soundtrack intact. This would allow the user to remain within the acoustic space of the programme but with a major source of masking removed. Here, atmosphere and effect sounds help contextualise the programme while the music conveys some of the scene's mood. Modifications to the MPC could be taken one step further by removing the music from the soundtrack, based on the argument that users may still find the music channels distracting. On the other hand, as the emotion conveyed by music is likely to be relevant mainly to the ongoing action in the MPC narrative, it will be less relevant to the SPC. Meanwhile, the continuation of atmosphere tracks keeps the listener immersed in the same space as the programme, while causing minimal disruption to the SPC. The degree to which such modifications would affect the user experience is unclear. Muting elements from the MPC to more successfully convey SPC may be perceived as intruding on the main programme. The user may also still feel that the removal of the key informational channels of speech and music breaks their immersion in the show. Furthermore, while full understanding of both streams in parallel seems unlikely, users may be inclined to switch attention on hearing elements from the speech or music (e.g., the exclamations noted by Brown *et al.* (2014)). The findings of Guerreiro & Gonçalves (2014, 2016) suggest that users are able to cope with concurrent streams of speech like this and to tune into elements of interest in purely auditory presentations. By removing these channels, the user is denied this opportunity. The degree to which a user would choose to do this and how these manipulations would affect the user experience is currently unknown. The effect of these different manipulations of MPC is a subject that needs further investigation.

As highlighted in Section 6.2, these soundtrack elements may also have some impact on the user-experience of VSPC. Through removing elements of the MPC soundtrack, it may be possible to reduce the amount of disruption caused by the unattended stream whilst maintaining cohesion between the primary and secondary components of the experience. One could envision the development of an attention-aware system that determined the locus of the user's visual attention and manipulated the auditory presentation to facilitate a better

experience of both streams of content. With the presentation of VSPC, user attention could be detected using gaze detection systems on the main screen or secondary device, and/or through detecting user-interaction with the secondary device.

### 6.4.3 Talker/performance factors

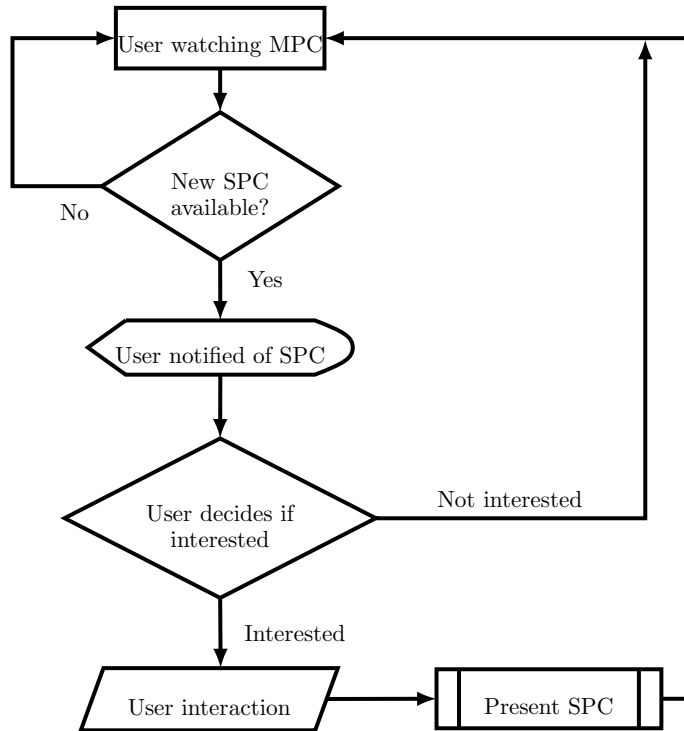
Adding speech to MPC requires that the characteristics of the additional speech are given some consideration. It is desirable that the user is able to distinguish SPC from MPC even in scenarios in which the speech is not overlapping. This is important to help the user to understand the model of two streams of information and to decide when to switch attention from one to the other. One manner in which this may be effectively communicated is through the identity of the voice that is speaking the SPC.

MPC may contain speech from characters within the scene or narration. Of these two types of speech, it is the narration that is most likely to cause confusion with spoken SPC. Characters are usually present in the visual MPC, though not necessarily for the duration of their speech. This allows the audience to attribute voices to different characters within the programme. Also, the voices of characters within the programme are generally mixed so that they sit within the acoustic space of the scene, while narration is generally recorded and processed to be clearly outside the acoustic of the scene, either through close-microphone recording in an acoustically dead room, or by applying considerable reverberation to give the effect of hearing a character's thoughts (Holman, 2010). The traits of the former treatment are shared by ASPC, which is why it is most important to focus on ensuring that these streams are separable. This may be achieved by ensuring that the voices of the narrator and the ASPC are qualitatively different.

The sex of the talker is an important factor in voice identity. Using talkers of different sexes has been found to be a factor in multi-talker scenarios (see discussion in Section 3.3.3). Ensuring that the ASPC and narrator voice are of different sexes is, therefore, likely to be particularly useful in scenarios in which both narration and ASPC are presented concurrently. In fact, even in the non-overlapping scenario of AD, it is recommended that the describer "be of the opposite gender to the narrator, to avoid confusion" (Incorporated Television Company (ITC), 2000, p. 8).

An issue is presented by the scenario in which both ASPC and AD are added to a programme. It is obviously no longer possible to rely on binary distinctions of sex to differentiate streams. In this scenario a talker should be chosen whose voice is qualitatively different from both the





*Figure 6.2: Flow diagram of proposed user interaction*

AD and the narrator, using factors such as voice pitch, accent, etc.

#### 6.4.4 Interaction design

Considering the potential presentation forms discussed in the last section, it is clear that presenting unwanted ASPC could be disruptive. With AD, the functionality of turning on or off descriptions is total; either AD is on and all descriptions will be included for the show or it is off, unless the setting is explicitly changed by the user. It is notable that this bimodality has generally been applied in VSPC experiences. This is not, however, a desirable interaction model for ASPC presentation. Interest in one piece of ASPC is not a reliable indicator of interest in all subsequent SPC. Such an approach would be particularly problematic using the ASPC presentation methods outlined in the previous section where the addition of ASPC may either obscure or remove elements of the SPC soundtrack. The user must be given the ability to make an informed choice on a case-by-case basis for the presentation of SPC. This may be achieved through providing a notification of SPC content prior to its display, to which the user may respond through a simple interaction to indicate that they wish the content to be presented (see Figure 6.2). In keeping with the non-visual nature of the display, it is desirable that the interaction should also not be reliant on the user's visual attention. This could take the form of a simple swiping gesture on the touchscreen of a portable device, a

keystroke on a semi-portable device, a button press on a remote control, or a single button on a special-purpose interface. It is recognised that the presence of notifications—the design of which will be outlined in the next section—may prove annoying for some users. It is proposed, therefore, that users would be given the ability to toggle the use of SPC and notifications in a similar manner to the addition of AD.

#### 6.4.5 Notification design

As was alluded to in the previous section, the system interaction requires the use of notifications to inform the user of new ASPC so that they may decide whether or not they wish it to be presented. The design of these notifications is an important consideration. The effectiveness of the notification will play an important part in the system's user experience. If they are either overly disruptive or not descriptive enough they could be more detrimental to the user experience than if basic ASPC switching were implemented in a similar manner to AD.

It is apparent that, in order to create a complete non-visual experience, the notifications themselves should not rely on the user's visual attention. Considering auditory approaches, the use of short spoken descriptions would be able to communicate both the presence of new SPC and some information regarding its content. In this scenario, however, it seems likely that spoken notifications would not be ideal due to the additional presentation time they would require. In this context, notification duration is important. If the notification delivers unwanted information, a longer notification means a larger amount of the MPC will be disrupted by its presence.

While non-speech cues are unsuitable for representing the detailed information that SPC may comprise, they do have a potential utility within these displays as notifications. Neate *et al.* (2015) demonstrated that non-speech cues can be used effectively to orient attention to new VSPC. It seems likely, therefore, that they would be effective for signalling the availability of new ASPC. The power of non-speech, however, extends beyond that of providing a simple alert. It is feasible that a non-speech method may be used to communicate (limited) information regarding the contents of the SPC. This would allow users to make informed decisions on what SPC they access. Non-speech cues are likely to be more appropriate for this task, as they can be made considerably shorter than spoken equivalents, reducing interference with the MPC.

Haptic codes from the secondary device are an alternative to non-speech auditory notification.

The use of a haptic display has the potential to be particularly effective in this context, as it does not interfere with a modality currently used for the presentation of the MPC. These notifications could take similar forms to the non-speech representations discussed above. Without the use of expensive additional hardware, the patterns would be likely to be dependent on simple rhythmic variations for communicating different notifications, similar to those used in smartphones. Unlike auditory notifications, however, haptic notifications are unable to orient attention to locations other than that of the haptic device without relying on a learnt association. If SPC were presented from sources other than the secondary device, such as a phone, this would mean that the user's attention would need to be orientated firstly to the secondary device and then to the locus of the SPC. This could be particularly confusing if more than one SPC source location were in use.

One may also consider a multi-modal approach in which both non-speech and haptic notifications are employed. This approach could allow spatial information to be added to the information communicated by the haptic display through the use of a shorter auditory cue. It is unclear how successfully these cues would integrate, as the use of two streams of information could have a negative impact. For the purposes of this project, the focus will be purely on the use of non-speech auditory notifications without haptic reinforcement.

As has been highlighted several times throughout this thesis, the requirements of learnability and distinguish-ability place limitations on the number of concepts that can be sensibly represented using non-speech notifications. For any use of such notifications, it is necessary to introduce the user to their forms and meanings. Within the television use case, the number of elements and the granularity of the information communicated by the non-speech about the SPC is likely to be dependent on the level of abstraction at which the elements are standardised. If non-speech notification were developed for individual programmes, one would be able to encode a large amount of information about specific SPC elements within the notifications, as there would only be a few different categories of SPC in use. The number of non-speech elements one could sensibly introduce and expect the user to remember within a single programme is clearly limited. The situation could be improved if broadcasters were to develop a standard non-speech grammar for these elements. Due to the extended time-scale of use, users would become familiar with the different sounds and it may be possible to introduce a larger number of more complex representations. Hence, the granularity of information provided by the non-speech audio would become finer.

The prospect of generally applied non-speech sounds introduces potential issues of stylistic

clashes between notification and MPC. For example, whilst a motif played on a harpsichord may be appropriate for a period drama, it could be jarring for a science-fiction programme. One solution to this problem would be to rely on rhythmic, or relative pitch cues to identify the nature of the SPC and allow timbre to be defined on a show-by-show or genre-specific basis. This could work by using instrumental earcons without specified timbre, or by using environmental sounds and manipulating them, as with morphocons (Parseihian & Katz, 2012). Familiarisation with this form of encoding could be more difficult for the user to acquire, as it is necessary to explicitly separate the melodic and rhythmic elements from timbre. Alternatively, if all broadcasters agreed to use a common system there is no technical reason why users should not be allowed to create their own libraries of notifications. In this scenario, users would be free to choose the representations that they find easiest to remember or least annoying. This could operate much in the same way that alert sounds are handled on modern smartphones, where the user can configure different ringtones for different types of alert (e.g., phone calls, alarms, text messages and social media messages).

The information conveyed by these messages could be no more than a simple classification (e.g., character profile, location, actor/personality profile or behind-the-scenes fact), which on their own would not provide sufficient information about its content for the user to determine if it was of interest. The timing of its presentation could, however, be used to infer additional relevant information. For example, if a notification were to play indicating a character profile as the programme shows a character entering the room, or after a character is mentioned, it is clear to whom the profile will refer. Care must be taken in this situation, where visual cues are chosen to indicate the notification's context, to ensure that the context is also clear to a viewer who is unable to visually attend to the screen. This may be through ensuring the notification is timed to follow a relevant piece of AD or dialogue that provides sufficient context. This approach, of using the timing of the notification, may be thought of as providing the auditory equivalent for film of a footnote marker in text. In this scenario, however, the footnote marker provides additional information on the contents of the footnote.

#### **6.4.6 Spatial location**

Conventions have been developed governing the spatial attributes of mixing for film and television. As a new construct, the spatial location of SPC is worthy of some consideration. ASPC does not correspond with a visible sound source in the same manner as on-screen speech or Foley. There is not, therefore, a particular location from which users should expect

it. This raises a question about where it should be spatially located. Most of the spoken material within programmes is presented from the centre of the spatial image, with the exception of some off-camera dialogue (Holman, 2010). To follow this convention, it may be argued that ASPC should also be presented from the front-centre of the mix. Using a different spatial location, however, could be a useful way of differentiating the streams to the user and providing an additional cue to the talker identity cues, discussed in Section 6.4.3. When the programme is watched with AD, the sex of the ASPC voice can no longer be a distinguishing factor and, therefore, the use of distinct spatial sources could be particularly useful here. Due to the role of spatial location in auditory stream formation (Bregman, 1990), one would expect that, for scenarios in which the user were presented with concurrent auditory streams, a spatial separation between the sources would facilitate improved experiences using the display. Greater spatial release from masking is observed when target speech is masked by other speech than when it is masked by spectrally similar noise (Freyman *et al.*, 1999, 2001). The effect of spatial separation is therefore likely to be particularly powerful when both of the streams consist of speech, as would be the case for the presentation of ASPC. This would suggest that ASPC should be laterally offset to reduce masking effects with MPC.

The presentation of ASPC for synchronised companion experiences, however, is atypical when compared to most work looking into spatial separation of auditory streams. Firstly, the degree to which a user would want the ASPC to seem like an element of the MPC or a distinct informational stream is uncertain, as both ASPC and MPC are contributing towards the experience of one programme. It is possible that users may prefer conditions that prioritise the cohesion between primary and secondary streams over the advantage offered by spatial separation. Secondly, MPC cannot be considered as a single source in space. It is, in fact, a collection of sources distributed within the spatial confines of the reproduction technology. Within a stereo mix this restricts locations of the sources to a horizontal plane between the speakers in front of the listener. In 5.1, sources may be positioned on a horizontal plane around the listener. Defining spatial separation when adding a channel to a complete mix is therefore difficult. It is, however, common that on-screen action (i.e., dialogue, narration and Foley) is concentrated towards the front-centre of the sound scene so as to correspond approximately with the locations of visual stimuli, while occasional off-screen action may be panned to other lateral locations, and ambience tracks are generally multichannel recordings that are spread across the output speakers (Holman, 2010). The most detrimental interactions with the ASPC are likely to be caused by the core sound components.

As this research is set within the context of a television viewing scenario, it is also necessary to consider the locus of visual attention. Cross-modal links have been found between auditory and visual spatial attention, which make irrelevant audio harder to ignore when it arrives from an attended visual location (Spence *et al.*, 2000). This effect would suggest that in a spatially separated display, the viewer would always be biased towards information originating from the location of the screen. This may mean that MPC dialogue or narration is harder to ignore when the ASPC is presented from another location, or that users choose to divert visual attention from the screen so as to ignore the MPC speech more effectively.

There may be some benefit from having the ASPC source (appear to) emanate from a tangible object within the same room as the user. Providing a visible source may help the user to form a stronger model of the spatial configuration of the presentation and aid them in knowing from where to expect additional content. Having this object as something that the user may physically hold also introduces the prospect of proprioception reinforcing this effect. Psychophysical indications have been discovered which suggest that proprioceptive cues can reduce the amount of spatial processing required in performing selective auditory attention (Simon-Dack & Teder-Sälejärvi, 2008). This could be exploited through using either the speakers of a secondary portable device or through a binaural display in which the location of the physical source is tracked and its movements applied to a virtual source.

So far, this discussion has focussed exclusively on the spatial separation between ASPC and both auditory and visual components of the MPC. In having the sources separate, however, a further question is raised about what their positions should be. It is not desirable to alter the spatial image of the main programme, as this would result in audio-visual spatial discrepancies. With a display presented over acoustically transparent headphones, it would be possible to maintain the MPC's binaural cues and hence its spatial image and to present ASPC to only one ear. This would mean that the ear free of ASPC would be unaffected by its presentation. As the MPC soundtrack encompasses the listener, however, acoustic conditions at the ear to which the ASPC is being presented could make attending to the SPC difficult. If, instead, the ASPC is presented binaurally, this therefore might prove beneficial to its intelligibility.

Employing spatial separation also raises the issue of which side the ASPC should be positioned. Researchers have noted asymmetric responses to speech (e.g., Kimura, 1961a; Bolia *et al.*, 2001; Sætrevik, 2012). While a small proportion of people have been found to exhibit a bias towards speech presented on the left, a right-ear advantage (REA) appears to be

more common (Kimura, 1961a; Wexler & Halwes, 1983). This is an important consideration for the spatial layout of the display, as it suggests that the hemisphere of presentation may influence a user's ability and predisposition to attend to the secondary stream. Furthermore, it is unclear whether the secondary content should be presented to the advantaged or the disadvantaged side. In the former case, the SPC might dominate the MPC stream. In the latter case, presenting it to the disadvantaged side could lead to more difficulty in attending to the content of the SPC stream. In the absence of stronger evidence and on the basis that ASPC is secondary content, we have taken the decision to present it to the generally less dominant side.

#### 6.4.7 Personal or shared experiences

One issue that was highlighted with methodologies that relied on pausing the MPC was their incompatibility with shared viewing scenarios. Thus far, however, no discussion has been given to how shared experiences may be facilitated with the ASPC presentations that have been proposed. In fact, the MPC modification approaches outlined here appear to be afflicted with the same problem, where all users would be affected by one user's choice to access SPC. This is only true, however, if all users are presented with the same audio feed. As the modification-based approaches considered do not involve modifying the time-scale of the MPC or its visual elements, the difficulty is simply overcome by conveying individual audio mixes to the different users.

A similar issue has been faced within the context of facilitating AD in shared viewing scenarios. Researchers from BBC Research and Development proposed a method for delivering alternative soundtracks to users' connected secondary devices via IP to accompany broadcast television programmes (Armstrong *et al.*, 2010). Jolly & Evans (2013) also mentioned this use case as a potential application for the BBC's Universal Control system. More recently, the Royal National Institute of Blind People (RNIB) have performed a study which assessed the user experience of AD delivered through a secondary device (Rai, 2015). The study found that participants, who were already users of AD, were generally positive about using the application.

For the provision of ASPC, a similar approach using headphone presentations may be used. One could imagine this service assuming several forms. The most easily achieved method for providing such an experience would be to send individual mixes of the content to headphones worn by each of the users. Rendering of the multiple mixes could be performed on either the

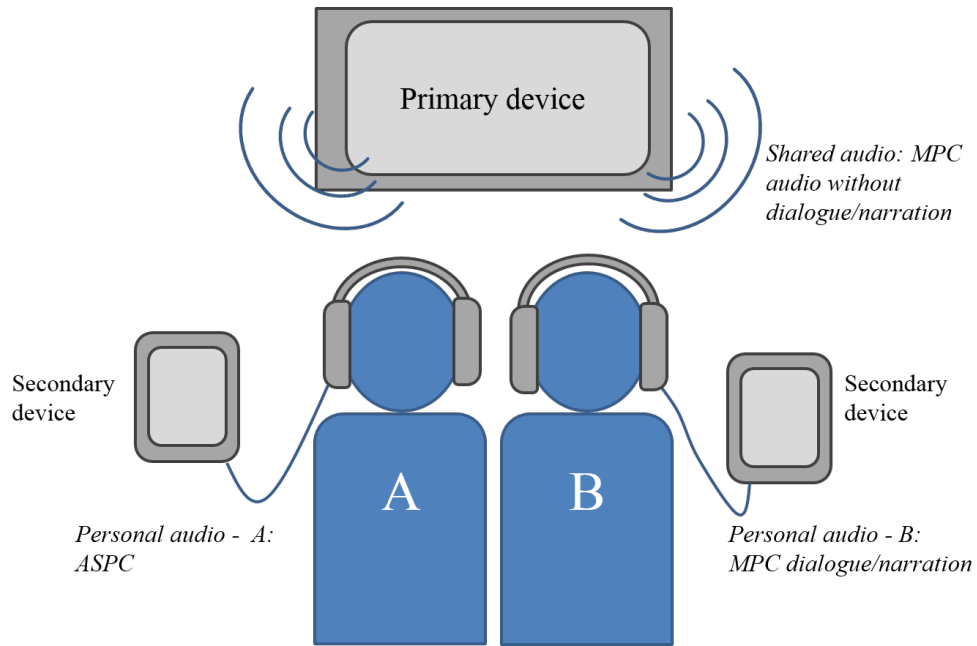
primary device or on the individual secondary devices connected to the users' headphones, as in (HbbTV Association, 2015). While the former reduces the amount of information that must be sent over the local network when the number of users is small, as the number of users increases, the amount of processing that must be performed on the primary device will be large and the amount of information sent over the network may exceed that of the unprocessed objects. Given the ever-increasing capabilities of portable devices and the progress that has been made on delivering synchronised media that has been reported by the HbbTV project for their current (2.0) specification (HbbTV Association, 2015), this seems a reasonable prospect for future display systems.

The ability to deliver spatial audio to users with binaural processing makes headphone presentation an attractive medium for providing users with engaging immersive experiences for television. Due to this, the use of binaural audio has attracted interest from broadcasters in recent years (e.g., BBC Research & Development, n.d.a; Pike & Melchior, 2013). Given the spatial auditory experiences this will allow, it seems likely that the use of headphones for the general consumption of television content can be expected to become more widespread across devices.

Wearing traditional headphones stifles communication between users and may be seen as being antisocial in a multi-user context. For this reason, it is important that any such headphone be transparent and accepted by the user. They would ideally not occlude the user ears, so as not to visually appear as a barrier to communication between users. A detailed consideration of the technical aspects of such a system is beyond the scope of this research, though the author imagines a technology akin to extra-aural (e.g., Erbes *et al.*, 2012) or bone-conduction headphones would be most suitable. The use of transparent headphones also introduces the possibility of augmenting traditional loudspeaker presentations with an additional spatialised channel for ASPC. If MPC were to remain unaltered, the presentation of ASPC would entail the trivial addition of presenting the ASPC via the headphones. This would necessitate the wearing of headphones only by users wanting to access the ASPC, whilst any number of others could consume the MPC on its own.

Scenarios in which the MPC is modified are also feasible. For these, the soundtrack may be split between shared and individual elements. Shared elements of the soundtrack may be delivered over the main system loudspeakers, while individual elements could be delivered via transparent headphones (see Figure 6.3). The user experience in this approach would be dependent on the proportion of the soundtrack deemed as shared and the fidelity of the





**Figure 6.3:** The use of shared and personal streams where the MPC speech is removed for the addition of ASPC. Here, person A has elected to access the SPC, while person B has not

headphones. Given the limitations associated with headphone presentations (e.g., typically a poor low-frequency response), this could be used to maximise the fidelity of the shared audio using the television speaker system. Where high quality headphones are employed, however, this hybrid approach would offer little benefit over the completely headphone-based approach discussed previously if the MPC soundtrack were modified.

At the time of writing, the technologies on which this vision are based do not exist in their ideal state and their development is clearly beyond the scope of this project. This research will focus instead upon the design principles of a single user's experience. This should be invariant whether the display is presented by transparent headphones with binaural processing to individual users or through a full loudspeaker set-up.

## 6.5 Secondary content applications

To help the reader appreciate the implications of the proposed design considerations, a number of imagined user journeys are provided. These have been specifically formulated to demonstrate the potential utility of SPC in serving a range of different users with different motivations for using auditory secondary content and to clarify particular abstract concepts covered in the design discussion.

### **Further interest in specific areas**

*Sue is watching a nature documentary enhanced with ASPC. During the programme, she is aware from the various notifications that different types of SPC are available. She generally chooses to listen to more information about the animals featured in the programme, occasionally she chooses to hear more about the location. She has little interest in the presenters or behind-the-scenes information, so usually ignores these notifications. When hearing a notification for information she is interested in, she performs a gesture on the secondary-device's screen to play the SPC. During the SPC, she continues to watch the MPC footage.*

### **A (mostly) shared experience**

*Andrea and Claude are watching a historical drama together. Andrea is really interested in the historical context of the programme, while Claude is interested only in the story of the programme. Andrea listens to the programme and an additional SPC track containing historical details. During the programme, they occasionally discuss the plot of the MPC or an interesting fact from the SPC. They enjoy the shared time, even though not all elements of their experiences were the same.*

### **An aid for memory**

*Dave is watching a long-running soap-opera. He watches everyday but struggles to remember all of the characters' complex back-stories, which means that he occasionally finds it difficult to understand the plot. When he thinks he is missing something, he listens to the character summaries to contextualise the action.*

## **6.6 Summary and outstanding questions**

This chapter introduces the concept of the auditory presentation of SPC as part of an orchestrated synchronous companion experience. Some similarities between AD and ASPC are identified and discussed. The work of I. Krejtz *et al.* (2012) and K. Krejtz *et al.* (2012) is highlighted as an interesting implementation which spans the two content types. Design considerations surrounding the presentation of ASPC are discussed, with reference to related work from researchers considering AD. The proposed ASPC display system notifies the user of

the presence of content with an informative notification which, if the user accepts it, triggers the presentation of ASPC.

While non-speech representations could allow the duration of ASPC to be reduced, the limited information they carry and the need for users to have to learn them mean that spoken representations are recommended. It is unclear how much concurrency between ASPC and MPC audio is acceptable. By applying concepts from object-based delivery of television programmes, it is feasible that programmes could be restructured to allow for the MPC soundtrack to be manipulated so that it facilitates the presentation of ASPC. A review of the literature on reading in the presence of auditory stimuli is presented, which suggests that a user must choose to process textual SPC or the MPC speech and that reading performance in the presence of the MPC audio could be reduced compared with reading in isolation. It is not clear that presenting VSPC concurrently with the MPC soundtrack is an optimal approach and, therefore, VSPC presentations may also benefit from the manipulation of MPC audio. Conversely, the fact that these factors are present within concurrent presentations of VSPC and considered to be acceptable user experiences, suggests that a similar level of interference between concurrent audio sources is acceptable. Further research is clearly needed to ascertain how these treatments affect the user experience of both auditory and visual SPC.

Non-speech is recommended for use as the notification. Design issues with these cues are likely to depend on the level of consistency that is reached across programmes and channels, due to the issues of learnability. Several different design methodologies for non-speech notifications are discussed. These different designs require user testing to determine how easily they are learnt, how effectively they communicate information about the SPC and how they affect a user's experience when engaging in an orchestrated synchronised companion experience. This is, however, likely to be dependent on the manner in which the ASPC is presented. As the correct way to display ASPC is not yet known, this is left as future work.

The spatial presentation of ASPC is discussed as a means of highlighting the distinction between ASPC, AD, and MPC dialogue. An emphasis is placed on configurations that do not affect the spatial image of the MPC soundtrack. Several different spatial configurations are suggested which either present the ASPC from the front, from the side, or using a handheld device.

The ability to provide both individual and shared experiences is highlighted as an important aspect for television user experiences. A discussion is provided regarding how this could be facilitated with ASPC presentations. Suggestions are made about the use of acoustically

transparent headphone-style devices that do not occlude the ear. The lack of such a device capable of providing suitably high fidelity audio, however, means that such a system is not realisable at this time. The thesis, therefore, is restricted to single-user scenarios.

The questions that have been raised over the course of this chapter require user testing. Firstly, it is unclear whether the MPC soundtrack should be modified to facilitate the addition of ASPC and, if so, what elements should be adjusted. Secondly, it is not known how the ASPC source would be best positioned. Furthermore, given the novel nature of these experiences, it is unclear how users will react to such presentations compared to the visual equivalent. These issues are therefore explored experimentally through a user study, which is the subject of the next chapter.

## Chapter 7

# The effects of secondary content modality, location and the modification of programme soundtrack on television user experience

### 7.1 Introduction

The previous chapter discussed design factors to be considered when presenting ASPC as part of an orchestrated synchronous companion experience. It also identified several aspects of the experience that require experimental evaluation to identify the best combination of factors. Questions raised specifically include the spatial configuration of ASPC and possible manipulations to the MPC's soundtrack. It noted that a comparison of ASPC to an equivalent VSPC experience would be valuable, and suggested that certain modifications to MPC audio may benefit text-based synchronous companion experiences. The chapter ended with several use cases of plausible scenarios involving SPC. These highlighted the factors of greatest relevance and most worthy of further investigation.

This chapter presents a methodology for investigating the impact of these factors on users' experiences, based on ratings of perceived disruption to specific elements, workload-style

questions, preference, and qualitative user comments. Results from this experiment are then presented and their implications discussed.

## 7.2 Experimental design

The system described in the previous chapter facilitates novel television user experiences. Though there are some similarities to visual synchronous companion experiences, it is likely that the evaluation of ASPC requires the consideration of new factors. This section identifies independent variables from the factors discussed in the last chapter. Possible ways of evaluating the user experience are discussed and dependent variables are selected.

### 7.2.1 Independent variables

The experiment reported in this chapter is concerned with two key factors: the presentation method used for the SPC (SPC source) and the treatment of the MPC audio (MPC treatment).

Chapter 6 identified centrally and laterally-positioned *fixed* locations for ASPC as being of primary interest. It was suggested that the laterally-positioned source should be located on the user's non-dominant side. It was also suggested that there could be potential benefits in having a tangible, *handheld* source for the ASPC. Based on this, and for the purposes of comparison, it was decided to include an experimental condition involving visual presentation of SPC from a handheld, portable device. This condition approximates the current norm in visual, orchestrated synchronous companion experiences. Hence, conditions of interest include the comparison of SPC as sound presented from  $0^\circ$  (*front*),  $90^\circ$  (*side*), using a handheld secondary device (*SD-A*). They also include the presentation of SPC as text from a handheld device (*SD-V*).

The previous chapter also discussed the potential impact that different elements from the MPC may have on the synchronous presentation of auditory and visual SPC. To explore this, several treatments of MPC are further included. These treatments are: leaving the soundtrack unmodified (*unmodified*), the removal of speech (*no-speech*), the removal of speech and music leaving atmospheric ambience (*atmosphere*), and muting (*mute*). It is acknowledged that lesser degrees of attenuation to specific elements, as opposed to muting, would be an alternative approach that may also prove beneficial. From an experimental perspective, however, the complete removal of the different soundtrack elements is likely to

exhibit larger, more easily observed effects and help to identify more clearly which elements of the soundtrack add value to the experience.

The effects of the MPC treatments may depend on the SPC source under consideration. For example, removing speech may be more beneficial when the ASPC is presented from the frontal location, compared to when the ASPC is presented from the side location. These factors require testing together, so that their interaction may be observed and interpreted.

### 7.2.2 Dependent variables

Unlike the experimental evaluation of menu design (Section 5.4.1), there are no clear performance metrics for these kinds of experiences. One approach could be to measure how much information is remembered after the experience, as in (Nandakumar & Murray, 2014). This, however, raises several experimental issues. Firstly, it suggests that the success of an experience is defined by how much it educates the user, and thus does not implicitly include other important experiential factors. Also, in a within-participant design, one must choose between taking measurements after each treatment or at the end of all treatments. The first of these options would soon make the participant aware that they are undertaking an attentional task and this might encourage them to alter their behaviour to try and remember information that they may be asked about. Thus, the experimental scenario would not be an accurate representation of normal television experience. The latter option is likely to be heavily influenced by ordering, which could lead to the effects from the factors of interest being obscured by those of nuisance variables.

Psychophysiological measures (e.g., pupilometry and galvanic skin response), offer an alternative objective quantitative approach and provide insight into emotional responses and workload throughout a treatment clip. These more intrusive measurement procedures can, however, negatively affect the ecological validity of the experiment and observed responses are subject to other factors in the environment and may be associated with other cognitive processes (Dirican & Göktürk, 2011). To avoid these issues, a method based on the subjective rating of elements of the user experience was chosen.

### Disruption

Adding information to a television programme introduces the potential for disruption to the experience. Disruption, in this context, may refer to energetic or informational masking

effects, attentional limitations and experiential aspects, such as a breaking the rhythm or mood of the MPC. With more information comes the possibility of increased workload for the user. Preference is an important indicator of how likely users are to make use of the experience under their own motivation. While disruption and workload are likely to play a substantial role in this, preference may also be affected by other factors.

Disruption is an important factor when considering the proposed experimental scenario. It may be caused by the effect of either of the two streams, MPC or SPC, on the other. It is therefore necessary to explore the disruption caused to and by each of these two informational streams. This has been considered previously by Neate *et al.* (2016), who presented a study on the visual complexity in orchestrated synchronous companion experiences, in which participants rated how challenging it was to consume content from each screen and how much they felt they missed material from each of the sources. A similar symmetrical approach is taken here. It is important to note that a unidirectional approach, in which only disruption to the MPC is observed, would suggest that an optimum system is one in which no disruption is caused to the MPC. This may include scenarios where the SPC is very difficult to attend to or is even absent. A symmetrical approach avoids these issues by seeking conditions in which the two streams cause minimal disruption to each other. A further complexity requiring consideration is that MPC contains both audio and visual streams of information, whilst the SPC used in this study is either entirely audio or mainly visual with a brief audio notification. It is feasible for disruption to occur between specific elements of the experience only. Furthermore, disruption to SPC may be caused by either the visual or auditory elements of the MPC. A single rating of disruption within the experience cannot separate these causes and may obfuscate important information. For this reason, we divide disruption into six elements to unpick the disruption caused by specific elements of the experience. Each element is formulated as a Likert-style rating on which participants mark the amount of disruption ranging from low to high. The rating criteria are:

1. Disruption caused by the secondary content to your experience of the main program
2. Disruption caused by the secondary content to your experience of the main program's soundtrack (i.e., voices, music, atmospheric sounds)
3. Disruption caused by the secondary content to your experience of the main program's visual content
4. Disruption caused by the main program to your experience of the secondary content



5. Disruption caused by the main program's soundtrack (i.e., voices, music, atmospheric sounds) to your experience of the secondary content
6. Disruption caused by the main program's visual content to your experience of the secondary content

The criteria are worded so as to place a focus on the user's experience of the elements, rather than on the content of them. This is important, because treatments may obscure elements of an individual stream without disrupting the user's experience of the stream as a whole. If the elements being obscured are not being attended to or are acting as a distraction to the user, their obfuscation may not disrupt the user's experience of that stream. By probing the effect of both modalities of the MPC individually, it is possible to provide insight into the subjective effects of inter- and intra-modal concurrency in the experience. This is particularly interesting within this experiment, as the modality of the SPC is likely to impact where the disruption is, as well as the total amount of disruption experienced.

### **Workload**

Five questions were adapted from the shorter versions of the NASA TLX questions, as featured in (NASA, n.d.):

- How mentally demanding was the experience?
- How physically demanding was the experience?
- How hurried or rushed was the pace of the experience?
- How hard did you have to work during the experience?
- How insecure, discouraged, irritated, stressed, and annoyed were you?

The modifications involve replacing the term "task" with experience. This is necessary, as the participants in the present experiment are not explicitly given a task to perform. A sixth question regarding perceived performance was removed. The lack of a specific task and the passive nature of the television viewing within the study do not fit well with the idea of performance. It was felt that this question would lead to confusion for participants and would not provide meaningful data.

## Preference

The final rating question is designed to get an insight into how much the user likes the display method. In the present experiment, the user experience does not include the interactive feature of being able to select to display the SPC (discussed further in section 7.3.2). It was deemed necessary to cue participants to imagine the interactive behaviour of the finished app. Due to practical constraints, this was considered the best way to get an impression of the participants' preferences without its measurement being overly disturbed by the lack of actual interaction in the study. The question was phrased:

- In an interactive version of this system, how would you feel about secondary content that you chose to access being presented in this way?

The question is scored on the same style of scale as the other questions, but with the labels changed to "I would not like it", and "I would like it". The wording of this question focuses the user on the display method and reduces the influence of the user's interest in the content itself during the trial.

## Comments

While the study has a predominantly quantitative focus, a qualitative element is included to provide support to the quantitative results and to identify other elements of the experience of importance to the users.

## Ear advantages and Handedness

In Chapter 6, ear advantages are identified as a factor that may be important in participants' use of displays when audio is presented from the side. While it would be interesting to investigate whether participants with REA differ from participants with left-ear advantage (LEA), this is beyond the scope of this project. To limit the effects on the experimental data due to this factor, the decision was taken to recruit only participants who identify themselves as right-handed. Handedness in itself is not a perfect predictor of the side on which a participant has an ear-advantage (Kimura, 1961a; Wexler & Halwes, 1983). Handedness has been found to exhibit a linear relationship with left-hemisphere language representation in the brain (Knecht *et al.*, 2000), which has been associated with REA (Kimura, 1961b, a), and right-handed individuals have been found to exhibit REA more consistently (Wexler

& Halwes, 1983). Whilst right-handed participants likely include participants with no asymmetry, or LEA, testing for this directly requires lengthy additional testing, deemed excessive for this experiment.

The 10 questions in the Edinburgh handedness inventory (EHI) (Oldfield, 1971) are used to quantify the handedness of the participants. The EHI requires that participants report which hand they prefer to use for common tasks. For each question, participants are asked to select: only left, left, no preference, right, only right, or don't know and these values are converted to left and right scores. The scoring method as used by Oldfield (1971) is shown in table 7.1.

**Table 7.1:** Scoring system, based on (Oldfield, 1971), used for calculating left and right scores for the laterality quotient.

Rating	Only left	Left	No preference	Right	Only right	Don't know
Left score	2	1	1	0	0	0
Right score	0	0	1	1	2	0

Right and left scores are used to calculate the laterality-quotient, as shown in equation 7.1 (Oldfield, 1971). Where  $R_p$  and  $L_p$  are the sums of the right and left scores, respectively, from all ten questions for participants  $p$ , and  $LQ_p$  is the laterality-quotient. The laterality-quotient ranges from -100 to 100 with negative values referring to left-handedness, positive values referring to right-handedness and greater magnitudes indicating a greater preference for that side.

$$LQ_p = 100 \times \frac{R_p - L_p}{R_p + L_p} \quad (7.1)$$

## 7.3 Methodology

Following on from the experimental design discussed in the Section 7.2, this section describes the methodology that was used to assess the different SPC source and MPC treatments.

### 7.3.1 Experiment structure

In this experiment, participants were presented with clips from television programmes with the added SPC under the various experimental conditions. Each clip that a participant was shown comprised one combination of the SPC source and MPC treatment conditions. Participants were then asked to provide ratings and comments after each clip.

The experiment was a mixed design, with SPC source as a between-participants factor and MPC treatment as a within-participant factor. SPC source had four levels [front, side, SD-V, SD-A], as did MPC manipulation [unmodified, no-speech, atmosphere, mute]. Participants were split into four groups. Each group was assigned a SPC source condition. This allowed the two main effects (SPC source and MPC treatment) and their interaction to be explored. A mixed design was chosen as a compromise between the duration of the experiment experienced by each participant and the number of participants that would be required in the study. The number of trials required was also a consideration, as original clips were needed for each one. Due to the novelty of the experiences under consideration, it was necessary to provide participants with some familiarisation material before the start of the experiment proper. It was decided that this would take the form of clips presented from a different programme encapsulating all of the SPC source conditions and the unmodified MPC treatment. This also served to provide all participants, regardless of their SPC source group, with an experience of all of the other ways that SPC would be presented across participants and to help inform their responses in the main experiment.

### 7.3.2 Experiential limitations

For the purposes of experimentation, it was necessary to make some simplifications to the suggested display features outlined in Section 6.4.

The availability of interaction to activate SPC elements, as suggested in Section 6.4.4, would allow participants to vary the timings of the SPC presentations or choose to completely disregard the SPC. Due to the experimental need to keep participant's experiences as similar as possible, it was decided to exclude the interactive element from the tested experience.

As participants would not be performing interactions to select which SPC elements were played, a decision was taken to use only one notification for all SPC elements within the experiment. This minimised the number of sounds participants would have to get used to in the experiment and removed the potential nuisance factor of which earcon was used in each case. The investigation of the benefits of specific design methods for notifications, as discussed in Section 6.4.5, is therefore left as an area for future work.

Although binaural technologies were discussed in the previous chapter as a possible means for content rendering, it was decided to use loudspeaker presentation for the prototype system. This decision was taken to remove any potential impacts from perceptual inconsistencies due

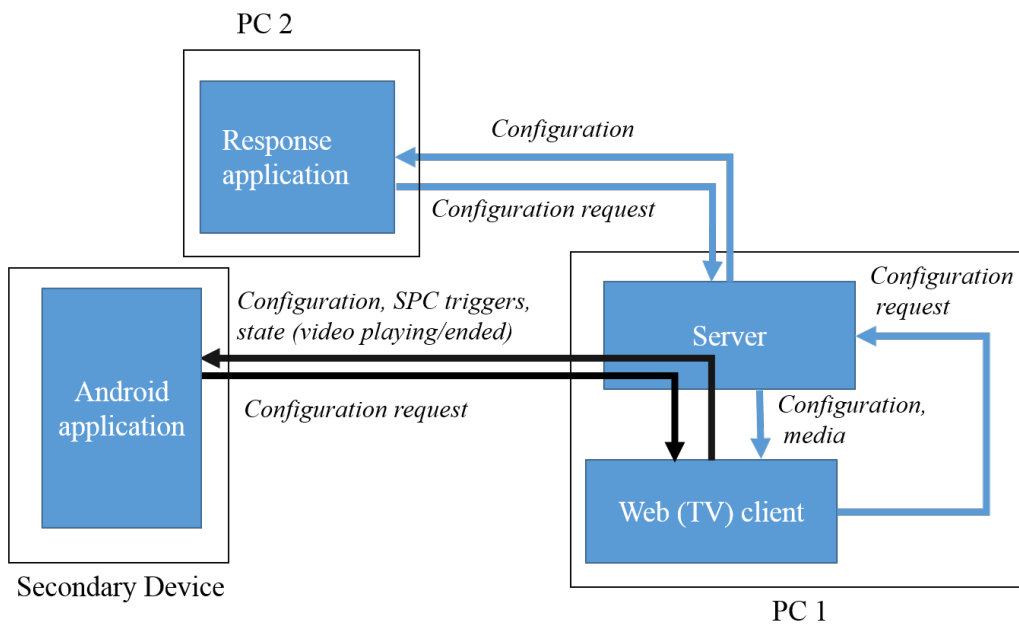
to the binaural processing, to reduce the technical complexity of the experiment, and to avoid any issues with headphone fidelity. From an experiential perspective, using a loudspeaker layout effectively represents the ideal binaural presentation without the associated overheads, such as capturing a full set of HRTFs for each participant and the use of a high-quality binaural rendering system. This approach does, however, differ from a binaural presentation in a few aspects. The user has visual cues corresponding to the location of the sources (i.e., the speakers) and there will be some influence from the acoustic of the listening space. The progress being made towards more advanced spatial audio systems for broadcast appears to have great potential. At the time of writing, however, television production primarily provides stereo or 5.1 mixes. It was decided, therefore, that the experiment would use surround sound content for MPC.

### 7.3.3 Prototype system

To run the experiment, it was necessary to create a system capable of providing the discussed SPC user experiences. It should be noted that, as the focus of this work was on the display principles, the technical implementation of the system was not intended to be representative of a fully developed system. This meant that the implementation could be considerably less complex.

As interaction was not designed into the system, simulating additional sources for the fixed ASPC conditions was performed by adding two channels containing the ASPC to the programme soundtracks and routing them to additional loudspeakers, positioned at the appropriate locations in the room. To simulate the *handheld* auditory and visual experiences, however, required a more complex solution. This was achieved using websocket connections over a local wireless network. The system consisted of a server, HTML5 page and an android application running on a Nexus 5 smartphone. In addition to the main experience elements of the experiment, a PC application was developed for the purposes of presenting instructions and gathering responses from participants in the experiment. This communicated with the server to gather information about the current experimental conditions and log them with the results in the database for analysis. This application will be henceforth referred to as the *response application*.

A HTML5 page was created to simulate the behaviour of the television or primary device using the HTML5 video element to present the MPC. This web page communicated with the Android application via the server to send configuration information. The page was also

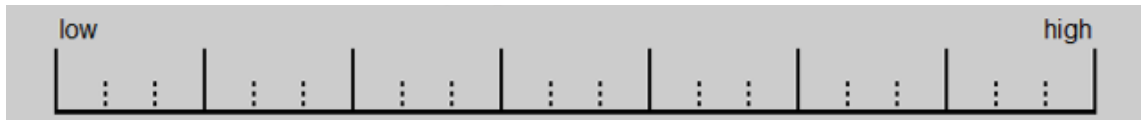


**Figure 7.1:** Diagram of the information passed between the individual elements of the experiment system

responsible for checking the play-head position of the video element and sending commands to the Android application when SPC should be triggered. The server was created using Node.js (Node.js Foundation, n.d.). The server was used to serve the HTML page and associated media, send configuration information to the response application and TV client on start-up, and pass messages between the web page and phone applications using a websocket connection (see Figure 7.1).

Initially, performing the MPC manipulation in the browser was considered, but keeping separate audio and video elements synchronised in the browser proved problematic. Though the tools have been proposed for this in the HTML5 specifications (W3C, 2014), at the time of development this had not been implemented in browsers. It was decided, therefore, to create different video files comprising alternative audio mixes for each of the MPC treatments. For the scenarios in which the ASPC originated from fixed locations, the WebAudio API was used to control the playback of the added ASPC channels and route them to the appropriate channel on the output device, depending on the experiment condition.

The Android application waited for messages from the web client before presenting SPC. In order to reduce the amount of information passing over the network and to simplify the system, all of the SPC media was stored locally on the device. During the playback of video on the primary device, the application displayed an empty white screen. When SPC was presented visually, it appeared as black text in a scrollable text box. This text was then



**Figure 7.2:** Example of the format used for the rating scales of the Likert-style questions in the response application

available for a set duration, corresponding to the same duration as the equivalent spoken SPC, before being removed.

The response application was built in Python and used Tkinter for the user interface. It presented participants with the rating questions and recorded their responses. Ratings were recorded on scales which were split into 21 sections comprising seven major sectors, each split into three sub-sectors (see Figure 7.2). The idea of this approach was to allow the user to more easily make meaningful ratings with a high resolution. By creating visual groups the user can first make a decision about the general region of the scale to use (i.e., which of the major sectors), followed by a more precise decision about which sub-sector to select. Similar scale designs can be found on some versions of the NASA TLX (e.g., NASA AMES Research Center, n.d.). It is acknowledged that the number of gradings within the scales used here is high for a subjective rating. A high resolution scale was chosen, however, to provide adequate granularity to capture both within- and between-participant effects in the  $4 \times 4$  design.

### 7.3.4 Stimuli

Two programmes with surround sound audio mixes were used for the experiment, one for familiarisation and a separate one for the experimental trials. The familiarisation clips were taken from an episode of “Upstairs Downstairs” (Jobst, 2012), which is a period drama set in Britain during 1938. The experimental stimuli were taken from an episode of “Lost Land of the Jaguar” (Backhouse *et al.*, 2008), which is a natural history documentary following a group of scientists as they survey the wildlife living in a rainforest in Guyana. Stimuli were taken from the original transport streams used for broadcast.

Points in the programmes were identified where SPC could be added. For the experimental material, care was taken to ensure that the soundtrack comprised some speech, music and atmosphere during the period in which SPC would be presented to allow the MPC manipulations to represent noticeable variations. For the familiarisation material, as the MPC soundtrack was not to be manipulated, this was of lesser importance. Due to the nature of the familiarisation programme, atmospheres were generally much more subtle than



*Figure 7.3: Notation of the secondary content notification earcon (tempo: 120 bpm)*

those from the natural history documentary. It was, however, ensured that sections where SPC was added comprised noticeable dialogue and music.

The transport streams were demultiplexed into forms that could be imported into a video edit suite. Here, the audio and video were manually resynchronised and then four clips were cut from each of the programmes. For the familiarisation programme, each clip contained one piece of SPC, whereas each of the experimental clips contained two. The stimuli used for familiarisation had durations ranging from 1m 15s to 1m 19s, while the experimental stimuli ranged from 3m 12s to 4m 14s.

### Creation of secondary programme content

Relevant SPC topics were identified for the selected points in the MPC. These included: profiles of characters/presenters, contextual information about the location or time period, more detailed information regarding something featured in a show (e.g., animal information and the importance of trees in global warming).

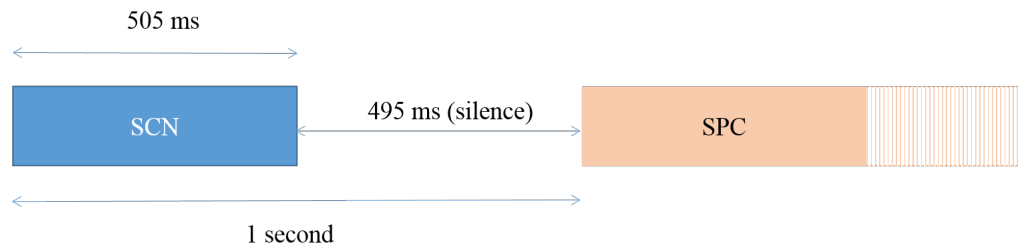
A list of facts was compiled for each SPC element, comprising information that was not otherwise covered in the clips. A freelancer then wrote passages using the information provided and recorded spoken versions. A female talker with a neutral British accent spoke the SPC. The freelancer delivered 24-bit 48 kHz audio files and the scripts used in the recording. The scripts were then used as the source for the visual SPC<sup>[1]</sup>.

The recordings were cut to reduce any excessive silences at the beginning or end of the files and a constant gain was applied to all of the recordings so that the clip with the maximum amplitude peaked at 0.9999. The duration of the recordings varied from 28.0 - 30.5 s.

The notification was an earcon that consisted of a marimba playing two notes (see Figure 7.3). It was generated using the Marimba pre-set of the EXS24 sampler in Logic Pro (8.0.2). The attack duration of the marimba was increased to  $\approx 15$  ms to soften the earcon's onset. This was done to reduce the intrusiveness of the earcon, while still allowing it to stand out from the MPC. A gradual fade-out ( $\approx 116$  ms) was applied to the earcon to shorten the

<sup>[1]</sup>Some small modifications were made to the scripts so that they matched the exact wording of the spoken content.





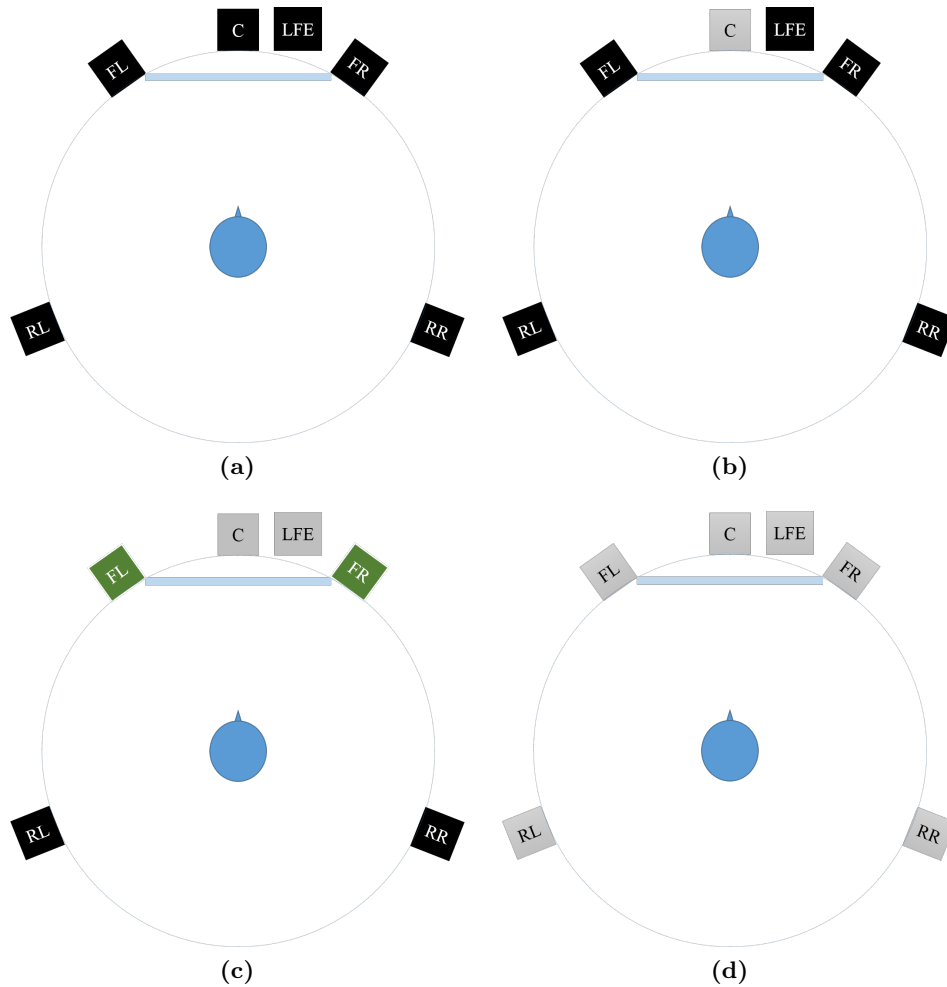
**Figure 7.4:** *Timing of the secondary content notification and secondary programme content*

decay.

The earcon was 505 ms in total duration and was exported with 495 ms of silence following it so that the ASPC could be appended to it and start one second after the start of the notification (see Figure 7.4). This gap between notification and SPC was to ensure that the notification and content were interpreted as two separate events and to more closely simulate how a display would sound that required interaction following the notification to activate the SPC. This also allowed time for fading down elements of the soundtrack after the notification in modified MPC treatments.

For the fixed ASPC conditions, the earcons and recordings were added to additional channels in the edit suite. Both earcon and ASPC levels were kept constant across all clips. The level of the earcon was set so as to be clearly audible, but not excessively loud. While the ASPC was set to be of similar loudness to the narrator in the experimental MPC.

In the *unmodified* condition, the MPC soundtrack was left unaltered and continued throughout the presentation of the SPC. For the *no-speech* condition, the centre channel of the soundtrack was removed. This also removed those effects which were panned to the centre channel and the atmosphere caught on the dialogue channels in the recording. In the *mute* condition all channels of the programme mix were silenced. In the *atmosphere* (or *atmos*) condition only the atmosphere and effects continued. To achieve this effect, a tape containing stereo mix-downs of different audio components from the programme's soundtrack was acquired from the BBC archive. This included a mix-down comprising the atmosphere and effects. Auditioning these elements alongside the programme in the edit suite uncovered synchronisation issues. It was therefore necessary to synchronise these elements with the programme soundtrack for each of the passages in which the SPC was added. To create the effect of all sound apart from the atmosphere and effects being removed when the SPC was displayed, the four front channels (including the subwoofer) of the soundtrack were removed and replaced by the stereo atmosphere and effects mix. The final condition was the soundtrack



**Figure 7.5:** The MPC manipulations used in the pilot: (a) unmodified, (b) no-speech, (c) atmosphere and (d) mute. Black boxes represent speakers presenting original MPC soundtrack, green boxes represent speakers presenting the atmosphere and effects mix, and the grey boxes represent muted channels.

as broadcast and, therefore, required no modification (see Figure 7.5).

### Main content modifications

Automation was added in the video edit suite to perform the modifications of the MPC. For the modified treatments, soundtrack elements that were to be removed began to be faded-out 13 frames ( $\approx 520$  ms) after the start of the notification and had been completely faded out by the start of the secondary content, 12 frames later ( $\approx 480$  ms). In the *atmosphere* condition, this approach was effectively reversed for the atmosphere and effects tracks, which began a fade up over 12 frames, 13 frames after the start of the secondary content notification.

Files were exported as QuickTime videos with eight discrete 48 kHz, 24-bit audio channels and then repackaged into MPEG-4 files using FFmpeg (FFmpeg Developers, n.d.) so as to

be compatible with Chrome. The libfdk-aac codec was used to convert the audio into AAC format with the low-pass filter at its highest possible setting of 20 kHz.

Both source videos were taken from high definition broadcast transport streams from the BBC. The videos tracks were broadcast at 25 frames-per-second (fps) and used h264 video coding. The video tracks of the two programmes were in different aspect ratios in the transport stream. The familiarisation programme was  $1440 \times 1920$  anamorphic, while the experimental programme was  $1920 \times 1080$  with square pixels. Clips were exported from the edit-suite to preserve the aspect ratio but converted to progressive video to avoid interlacing artefacts.

### 7.3.5 Pilot Study

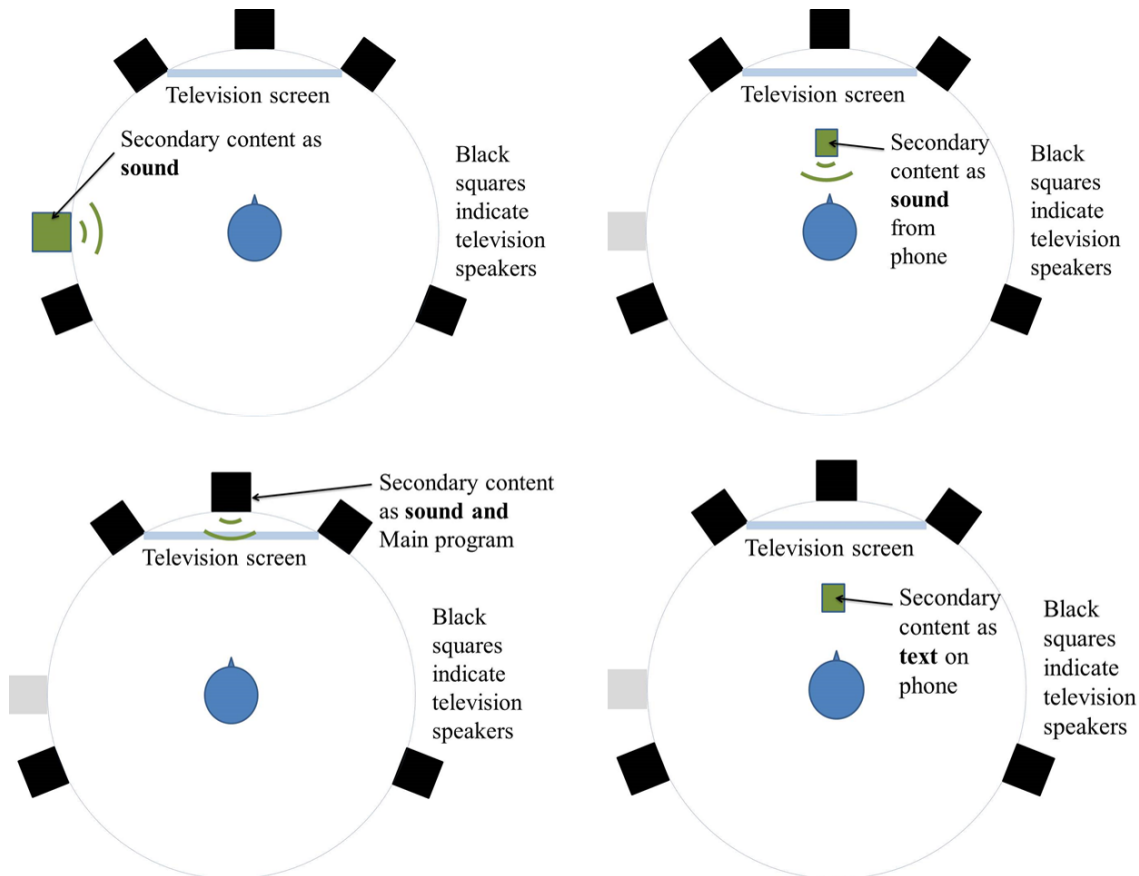
In order to test the experimental set-up, stimuli and methodology, a small pilot study was run prior to the experiment proper. The study comprised eight participants, with two participants assigned to each SPC source condition. Participants were recruited from BBC Research and Development.

#### Procedure

Participants were asked to complete an informed consent form, and then enter demographic information (age and gender) and completed the handedness questionnaire using the response application. They were then presented with written instructions explaining the familiarisation section of the experiment.

During familiarisation, participants were given a paper copy of the rating scales from the experimental section and asked to consider them while they watched all four familiarisation clips. Before each of the clips, an image was displayed indicating the location of the SPC in the following clip (see Figure 7.6). Irrespective of their group, all participants watched the same sequence of familiarisation videos with the same treatments.

Following the familiarisation, participants were presented with a second set of instructions explaining that all remaining trials would have the SPC in the same location. To avoid doubt, the corresponding image (as in Figure 7.6) indicating the location of the SPC was presented before each clip. Participants then watched each of the four experimental clips with different MPC treatments. As the pilot only involved eight participants, the two participants assigned to each SPC source group were presented with different orders of MPC treatments.



**Figure 7.6:** The four images shown to participants before trials to indicate how the SPC would be presented.

For both familiarisation and experimental sessions, all participants saw clips in the same sequence, which corresponded to the order in which they appeared in the original programme. As the clips were given a particular order for editorial reasons during the creation of the programme, it was felt that randomising the order of the clips could create confounding effects in the experimental data.

In the instructions, participants were informed that the finished system would be interactive:

“This experiment is intended to simulate an interactive system in which a short audio notification informs you that some new secondary content is available. In actual use, if you wished to access the content you would then perform a simple action to display it. In this experiment, however, the system is not interactive and secondary content will be automatically displayed following the notification sound. The notification and the secondary content will be presented from the same location with a short delay to separate them.”

After each clip, participants were given the laptop running the response application and

asked to respond to the rating scale questions. A free-response text box was provided for any additional comments.

After running through the experimental session, there was a brief unstructured interview with the participants to discuss their experience of the experiment including factors such as the understanding of the questions and how the stimuli were presented. This was intended to highlight any potential misunderstandings and address issues before the main experiment. The sessions were also video recorded to observe any interesting behaviours.

### **Experiment set-up**

A Mac mini was used to run the server and the TV client using a 47" television to display the video. The Mac mini output audio via a Focusrite Pro 40 interface. The six channels of the 5.1 mix were output through a Bluesky bass management system which then routed the five channels surround channels (i.e., left, right, centre, surround left, surround right) to individual PMC DB1S-A's, while the low frequency effects channel was routed to a PMC LTE1. The two additional ASPC channels were sent directly from the Focusrite Pro 40 to PMC DB1-A's.

The surround channels were arranged in a 5.1 configuration with a distance of 2 metres from the listening position. The centre channel speaker was positioned on its side above the screen with the central SPC channel placed on top of it. Positioning the speakers on their sides served to reduce the elevation mismatch between the audio and on-screen features for the MPC and minimise the elevation difference between the central and side SPC positions. The other SPC channel was positioned 90° to the left of the listening position, at the same elevation as the MPC surround speakers.

The Mac mini was connected to a Linksys E4200 wireless router with a private network which allowed messages to be passed between the server, the secondary-device (Nexus 5) and the laptop running the response application (Dell Latitude E7450).

### **Findings**

In the first session, the participant pressed a button on the secondary device to exit the application, despite the comment in the instructions to avoid doing this. As a result of this, subsequent participants were shown the buttons and verbally instructed to avoid pressing them during the experiment.

Initially, the phone was left on a coffee table for the user for them to pick up. During the pilot, however, it was felt that participants seemed reluctant to pick up the secondary device and engage with it. This was attributed to the participants not feeling any ownership of the phone. To counteract this, it was decided to hand the phone to the participants before each presentation and then allow them to position it in whichever way they chose.

It was identified that having all of the disruption questions on one page could lead to some confusion, as half referred to disruption of MPC on SPC while the other half referred to disruption of SPC on MPC. Furthermore, it was noted that the preference question scale runs from “I would not like it” on the left to “I would like it” on the right, meaning that positive ratings are in the rightmost sections of the scale. This contrasts with the rest of the ratings which measure negative factors (i.e., disruption and workload) with low ratings to the left and high ratings to the right. This was identified as a factor that could lead to participants accidentally providing inaccurate ratings.

Also, from discussing the preference question, it was found that participants differed in their understanding of what an interactive version would comprise and whether they considered the interactive system when providing a response. This was taken to suggest that the description in the instructions had not been sufficient to encourage participants to imagine the interaction involved in the final system.

The pilot also uncovered a number of technical issues which needed to be remedied for the experiment proper. Firstly, some issues with audio-visual sync were noted for the main programme content, which was attributed to a software conflict between audio applications running on the computer. Secondly, the pilot trials ran over two days and it appeared that the level settings were lost as a result of restarting the system. This led to differences in the configuration between participants that completed the pilot on each of the days. It was determined that this issue was easily avoidable through proper recording and re-entering values on start-up.

During the pilot it was also felt that the level of the MPC stimuli used for the training was slightly lower than the experimental stimuli. This led to the SPC appearing louder in relation to the MPC during the familiarisation than in the experimental sessions. It was therefore, determined that some modification of the programme levels would be required prior to the experiment proper.

### 7.3.6 Methodological changes following the Pilot

After the pilot study, some changes were made to the stimuli and system to counter potential issues that were identified. Small modifications were made to the stimuli to refine the timing of SPC. In addition to this, programme loudness was analysed using the “ebur128” filter in FFmpeg (FFmpeg Developers, n.d.) on the programme audio, which was trimmed to remove continuity speech from the transport stream. The programme levels differed by 5.8 loudness units referenced to full scale (LUFS) (also known as loudness, K weighted, relative to nominal full scale (LKFS)—defined in (ITU-R, 2011)). A gain of 5.8 dB was therefore added to the clips from Upstairs Downstairs to compensate for the difference. Clips were checked to ensure that no clipping occurred due to this change prior to export. Again, levels of the ASPC were set for the fixed locations to be a similar loudness as the narration in the experimental MPC, while the notification was set to a level so as to be apparent but not excessively loud. None of the changes that were made to the stimuli affected the duration of the clips for the familiarisation or experimental trials.

The way in which the ‘atmosphere’ MPC treatment was achieved was also altered. Rather than presenting the stereo atmosphere and effects mix from the front stereo pair and leaving the surround channels from the original soundtrack, the stereo atmosphere and effects mix was upmixed to 5.1. Up-mixing a stereo track to 5.1 is a compromise and is not equivalent to performing a full 5.1 mix on the original source materials. Given the lack of clean 5.1 material, however, this approach could not be avoided. The up-mix was performed using an experimental algorithm developed in BBC Research & Development that up-mixed using a simple linear combination of the original stereo channels (details in Appendix C) (Marston, 2016).

In response to the feedback on the rating questions, it was decided to split questions referring to the disruption of the SPC to the MPC onto a separate page to the questions about the disruption of the MPC to the SPC. It was hoped that this separation would help participants distinguish between the two sets of questions. The preference scale was also moved onto its own page in the response application due to the discrepancy between the directionality of the scale for preference compared to the other factors. While it would have been possible to have reversed the scale (i.e., running from ‘I would like it’ to ‘I would not like it’), it was felt that this would have seemed unusual and may have led to further confounding effects. By putting it on a different page, it was hoped that this difference would be less problematic. Screenshots of the response application are included in Appendix D.7.

## 7.4 Experiment proper

### 7.4.1 Demographics

In total, 35 participants from the University of York took part in the study, 32 of whom were included within the statistical analysis with eight assigned to each SPC source group. Participants were recruited through emails sent to university mailing lists, posters distributed across campus or were approached in person. Participants were offered a payment of £5 for taking part in the experiment. Participants' ages ranged between 18 and 45 with a mean of 22.5. All participants reported that they considered themselves to be native English speakers, right-handed, have normal hearing, and normal or corrected to normal vision. The sample contained 17 male (3 of whom were not included within the quantitative analysis due to the procedural anomalies and incorrect assignment discussed further in Section 7.4.4) and 18 female participants. The participants were randomly assigned into groups based on the order in which they did the experiment. As a result of this, groups did not have equal male-female ratios. All of the groups used within the quantitative analysis had eight participants. Out of these groups the visual SPC had the lowest number of male participants with two, and the secondary-device ASPC condition had the fewest female participants with three.

Analysis of the results of the Edinburgh Handedness Inventory of the full sample (calculated according to the description in Section 7.2.2) showed that all participants had laterality-quotients of 50 or above with the majority of the participants (19) having a laterality-quotient of 100.

### 7.4.2 Procedure

The experiment proper followed a very similar procedure to the pilot study outlined in the Section 7.3.5. Participants completed informed consent forms before entering demographic and handedness information into the response application. After this, participants were presented with the instructions for the familiarisation section of the study. Following the pilot study, it was decided to encourage them to act out the interaction during the trials by adding the following to the instructions (see Appendix D for full documentation):

“When you hear the notification, imagine that you perform a sideways swipe on the screen of the secondary device (i.e., the mobile phone) to trigger the secondary content. If you wish, you may find it useful to physically enact this gesture on



hearing the notification during the experiment.”

It was hoped that this would help to give them an impression of what the interactive system would feel like to use and assist them in providing a preference rating with the interaction in mind.

The familiarisation and experimental sessions were performed in the same manner as in the pilot. The one exception was that participants were handed the secondary device at the beginning of each clip if they did not already have it or rested it upon themselves.

The ordering of the MPC treatments was varied using two counter-balanced Latin squares. This ordering meant that all participants within a group experienced a different ordering of MPC manipulations. The order of the clips for the familiarisation and experimental trials was kept the same as the pilot—to correspond with the order that the clips appeared in the original programme. As with the pilot, participants provided ratings of disruption, workload and preference, and comments for each treatment they experienced in the experimental trials. The sessions were also filmed so that they could be reviewed and any interesting participant behaviours to be noted. After the final experimental clip, participants were debriefed and paid.

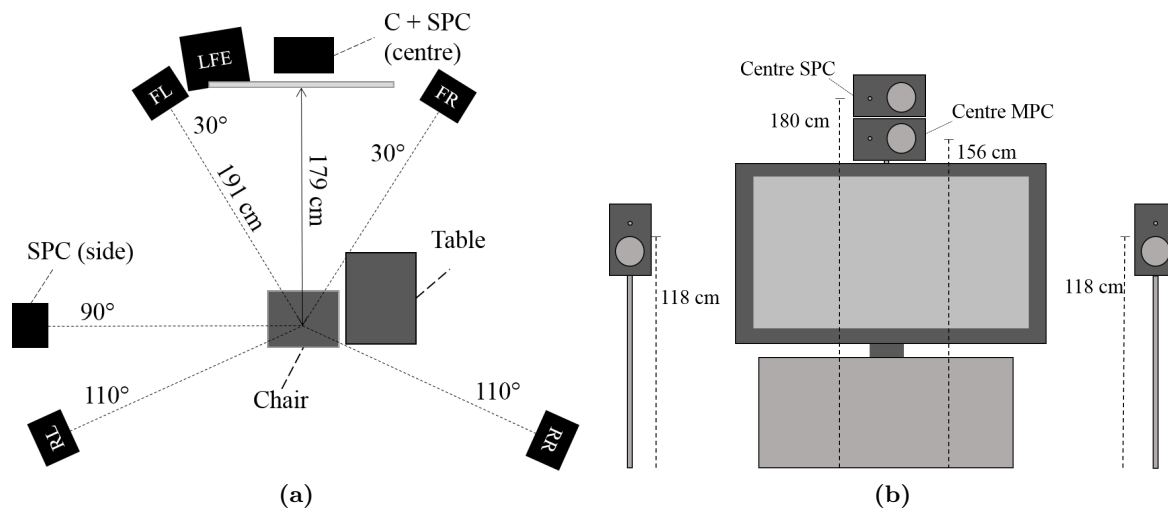
### 7.4.3 Experiment set-up

A Mac mini was used to run the server and the TV client using a 50” Panasonic TX-P50ST50B television to display the video. The Mac mini output audio via a Focusrite Pro 40 interface. The six channels of the 5.1 mix were output to a Genelec 7060B subwoofer with inbuilt bass management which then routed the five channels surround channels (i.e., left, right, centre, surround left, surround right) to individual Genelec 8040a’s. The two additional ASPC channels were sent directly from the Focusrite Pro 40 to Genelec 8040a’s.

The loudspeaker set-up consisted of a 5.1 rig, configured with a distance of 191 cm from the listening position, with the front stereo pair at  $\pm 30^\circ$  and the rear speakers at  $\pm 110^\circ$  (see Figure 7.7). While the configuration was intended to conform to the standard outlined in BS.775-3 (ITU-R, 2012), there were a few notable departures from this. Firstly, as in the pilot, the centre speaker was positioned above the television screen (156 cm<sup>[2]</sup>) on its side to reduce the elevation mismatch between the apparent location of the on-screen action

---

<sup>[2]</sup>all measurements refer to the approximate distance from the floor to the centre of the speaker unit



**Figure 7.7:** Diagrams of experimental set up (not to scale) showing: (a) a birds-eye view; and (b) a front view showing the heights of the speakers

and the associated sound. The elevation of the side and rear speakers was set at 118 cm. This was chosen as a compromise between height of the centre channel and the listener ear height (estimated as 112 cm). The television screen was positioned at a viewing distance of 179 cm. This is slightly below the ideal distance of 189 cm required to correspond with the recommendation of three times the height of the screen for HD screens in (ITU-R, 2012), but was a necessary compromise to fit the screen within the 5.1 speaker rig. As in the pilot study, two additional speakers were incorporated for the ASPC. One of these was positioned to the left of the listening position at 90° at the same elevation as the surround speakers while the other was positioned on its side on top of the centre channel speaker for the 5.1 system, resulting in a height of 180 cm.

It was necessary to fix the playback level of the programmes for all participants to maintain a constant relationship between the levels of the main and secondary content when the secondary content was presented as audio from the secondary device. All of the fixed speakers (5.1 and ASPC front and side), were calibrated to the same level at the listener position using pink noise. The system loudness was then set according to the dialogue loudness. A section in which the narrator’s voice was presented from the centre channel was found in the experimental material and the loudness from this speaker was monitored at the listener position. Originally it was intended to set this to 64.8 dBA, to correspond to Benjamin’s (2004) reported preferred dialogue level for television systems with external amplification and loudspeakers. This was, however, felt to be excessively loud. Instead, the level was set to 57.2 dBA to correspond approximately to the preferred dialogue level for televisions without



*Figure 7.8: Panorama of the set-up used for the experiment*

external speakers and amplification of 57.7 dBA as reported by Benjamin (2004).

As with the pilot study, the Mac mini, secondary-device (Nexus 5) and laptop running the response application (Dell Latitude E7450) were all connected to the same network from a Linksys E4200 wireless router.

A small unobtrusive camera (a GoPro) was positioned to the right of the screen (from the viewer's perspective - visible in Figure 7.8 between the screen and the front-right speaker) and used to record the session.

#### **7.4.4 Data treatment**

In completing the experiment, several technical issues were encountered. Also, the dataset's characteristics meant that the methods used for statistical analysis needed special consideration. This section provides details of the technical issues and subsequent treatment of the data from affected trials. Following this, the statistical methods used in the analysis of the quantitative elements of the dataset are presented.

##### **Technical artefacts**

The researcher sat at the back of the room during all of the experimental trials. A few small technical artefacts were noticed. These were not considered significant enough to disregard data from the experiment, but they are included here in the interests of completeness. The experimental nature of the prototype meant that there were a few occasions when small visual

or auditory glitches occurred. In a few of the trials the video appeared to periodically drop frames, leading to a slight jerkiness on panning shots. A couple of short audio drop-outs were also noted. It is not clear what caused these and they were not replicable. The notification from the secondary device appeared to stutter slightly occasionally. It is believed that this is due to an issue with the application losing ownership of the audio thread. Also, though efforts were made to move the cursor off the screen during the trials, parts of the cursor may still have been visible in all trials. Furthermore, there were a couple of trials in which the cursor was mistakenly left on screen. None of these artefacts was felt to impact the SPC presentations significantly during the experimental sessions.

Two participants were affected by procedural anomalies that could have impacted their experiences, which meant that additional participants were recruited in their place <sup>[3]</sup>. An additional participant was also required due to two participants being accidentally assigned the same participant number, and consequently being allocated the same condition ordering. The participants who experienced the anomalies during the experiment were excluded from the statistical analyses of the rating scales. Where two participants had been assigned to the same sequence of conditions, one was randomly selected for inclusion within the statistical analysis. As far as the qualitative feedback was concerned, only the clips that were directly affected were removed from analysis.

### Statistical methods

On inspection of the dependent variables collected in the experiment, it became clear that assumptions of normality and homogeneity of variance would not be valid.

Field (2009) recommends the use of the robust methods put forward in (Wilcox, 2012) when parametric assumptions are not met. Wilcox's (2012) methods are based on the idea of using measures other than the mean to characterise a distribution. It was therefore decided to perform the analysis based on trimmed means. As trimmed-mean analysis is not commonplace within this research field, a brief explanation is provided.

The trimmed mean refers to the mean value of the observations when a specified percentage of both the highest and lowest observations rounded down to the nearest integer value are excluded from the calculation. It follows that the arithmetic mean is a special case of the

---

<sup>[3]</sup>The television had an automatic time-out feature activated, which caused a large message to appear on the screen and the screen to turn off during one of the experimental clips for one of the participants. The other participant had an experiment clip interrupted when someone entered the room and in a different clip the controls of the video player repeatedly reappeared

trimmed mean with the trimming percentage 0% and the median is the trimmed mean with the trimming percentage of 50% (Wilcox, 2012). Wilcox (2012) refers to the 20% trimmed mean as a good measure for most scenarios. This value was used in this study.

Analysis of main effects and interaction was performed with a mixed-ANOVA with one between- and one within-participant factor, modified for trimmed-means (as described in (Wilcox, 2012, p. 408)). Significant main effects were further explored using pairwise comparisons. Significant main effects for SPC source were investigated using pairwise comparisons of the trimmed means (as described in (Wilcox, 2012, p. 317)), aggregated over the MPC treatments. Significant main effects for the MPC treatment factor were investigated with pairwise comparisons of the difference scores aggregated over SPC source treatments. This approach calculated the difference between each participant's ratings for each level and then tested whether the trimmed mean of all of the participants' difference scores significantly differed from 0 (i.e., it evaluated the difference between the ratings for the two treatments), as described in (Wilcox, 2012, p. 420). Though the arithmetic mean would be equivalent in these two approaches, this is not the case with the trimmed-mean. Comparisons of the two distributions' trimmed means excludes participants whose ratings are extremely high or low. Comparing the differences after trimming excludes participants whose trend in rating differed most from the rest of the group. The difference score, therefore, appears to be a better predictor for how an individual's ratings would change as a function of the different treatments. To control the family-wise error for the multiple comparisons, Rom's (1990) methods were used for the main effects and Hochberg's (1988) methods were used for the interactions.

Analysis was carried out in R (R Core Team, 2016) using the the functions in (Wilcox, 2016). The functions report p-values and provide critical values based on the appropriate adjustment of alpha for the multiple comparison procedures. To improve the readability of the results, the p-values have been adjusted to correspond to a critical value of .050 dependent on the correction method used, as described in (García *et al.*, 2010).

#### 7.4.5 Quantitative Results

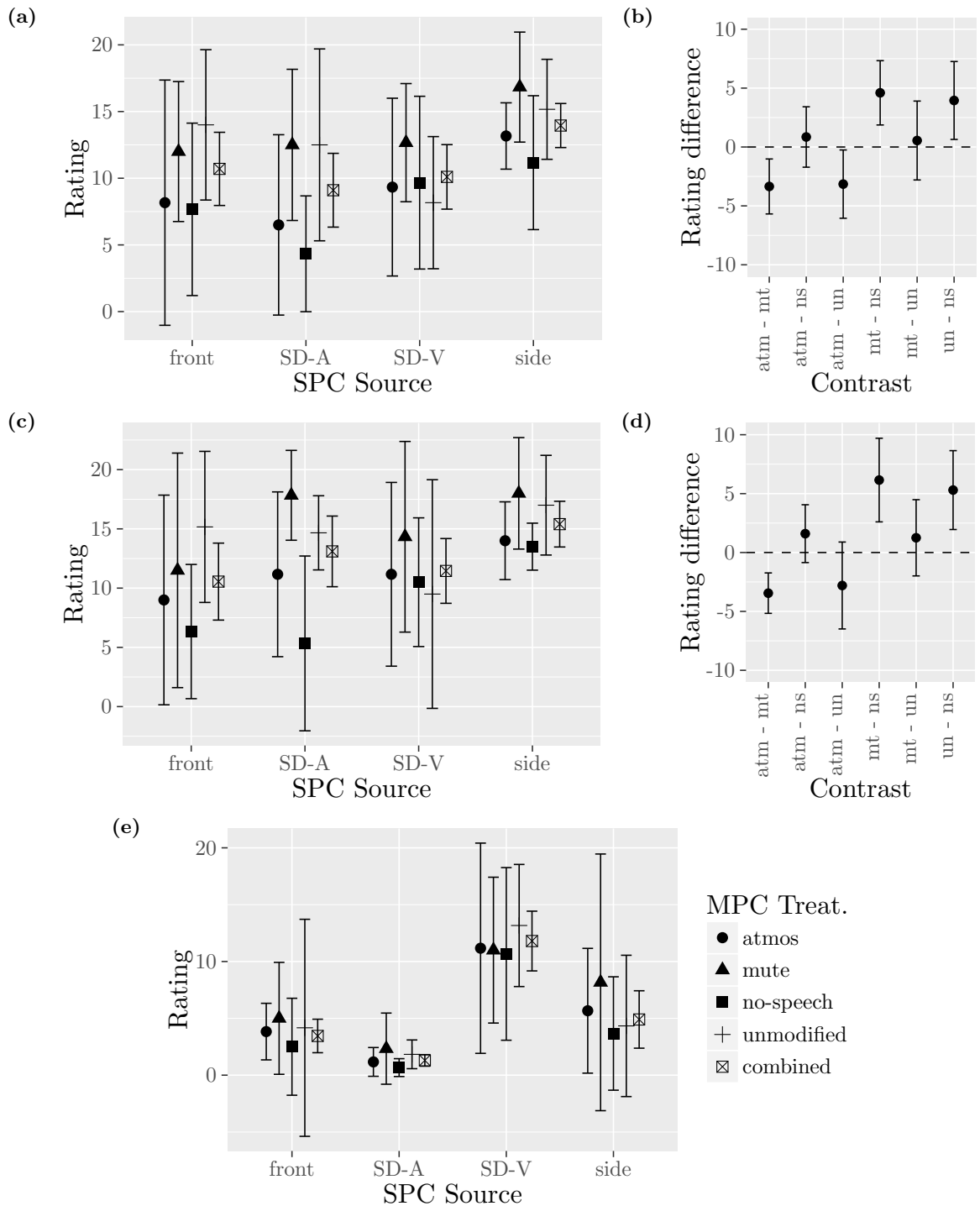
This section presents the results of the statistical analyses of the participants' ratings of disruption, workload and preference that were provided in the response application following each clip.

## Disruption

The ratings of disruption for the SPC on the MPC (Figures 7.9a and 7.9b) exhibit significant main effects for both the SPC source ( $F(3, 10.759) = 6.229, p = .010$ ) and the MPC treatment ( $F(3, 13.606) = 14.295, p < .001$ ), but no significant interaction between these factors is observed ( $F(9, 11.903) = 2.339, p = .086$ ). Pairwise comparisons for the SPC source indicate that the *side* group assigned significantly higher scores for the amount of disruption caused by the SPC than the *SD-A* ( $t(19) = 3.250, p = .016$ ) and *SD-V* ( $t(19) = 2.839, p = .036$ ) groups did. Pairwise comparisons for the MPC treatment indicate that the atmosphere condition was rated significantly lower than the mute ( $t(19) = -3.005, p = .028$ ) and unmodified ( $t(19) = -2.644, p = .048$ ) conditions, and that the no-speech condition was rated significantly lower than both the mute ( $t(19) = 4.414, p = .002$ ) and the unmodified conditions ( $t(19) = 3.13, p = .026$ ). All other comparisons are not significant.

The ratings of the disruption caused by the SPC to the MPC's audio exhibit significant effects for both factors (SPC source:  $F(3, 10.305) = 4.612, p = .027$ ); MPC treatment:  $F(3, 13.088) = 9.779, p = .001$ ), while no significant interaction is observed ( $F(9, 11.576) = 1.649, p = .210$ ). Pairwise comparisons of the differences for the MPC treatment indicate that ratings were significantly higher for mute than atmosphere ( $t(19) = -4.209, p = .002$ ). The no-speech treatment was rated as significantly less disruptive to the MPC audio than both the unmodified ( $t(19) = 4.355, p = .002$ ) and muted ( $t(19) = 4.766, p = .001$ ) conditions. Pairwise comparisons of the ratings pooled by SPC source reveal no significant differences between the treatments, though differences approach significance for the comparisons of the side and the front ( $t(19) = 2.771, p = .053$ ) and, to a lesser extent, the SD-V condition ( $t(19) = 2.54, p = .073$ ). Visual inspection of the data (see Figures 7.9c and 7.9d) shows that the trimmed mean ratings for all MPC treatments are higher when the SPC was presented from the left than when it was presented from the other locations. The one exception to this trend is the mute condition when presented as audio from the secondary device, which appears practically equivalent to the same MPC treatment with the SPC source originating from the left.

Disruption to the MPC visuals by the SPC (Figure 7.9e) displays a significant main effect from the SPC source ( $F(3, 8.980) = 10.028, p = .003$ ), but there is no significant effect from the MPC treatment ( $F(3, 12.059) = 2.587, p = .101$ ) or from the two factors' interaction ( $F(9, 11.172) = 0.684, p > .711$ ). Pairwise comparison of the SPC source conditions across the MPC treatments indicates that all groups' ratings differ significantly from each other,



**Figure 7.9:** Plots of the ratings of disruption for the SPC on the experience of: (a)(b) the MPC; (c)(d) the MPC audio, and (e) the MPC visuals. (a), (c) and (e) show 20% trimmed means for each MPC treatment in each of the SPC source groups with 95% confidence intervals (CIs). (b) and (d) show: 20% trimmed means of participant difference scores between the MPC treatments (Rom's corrected 95% CIs). X-axis labels refer to the MPC treatments being compared (e.g., atm-mt is the score from the atmosphere condition minus the score from the muted condition for each participant). If CIs do not cross zero, the difference is significant.

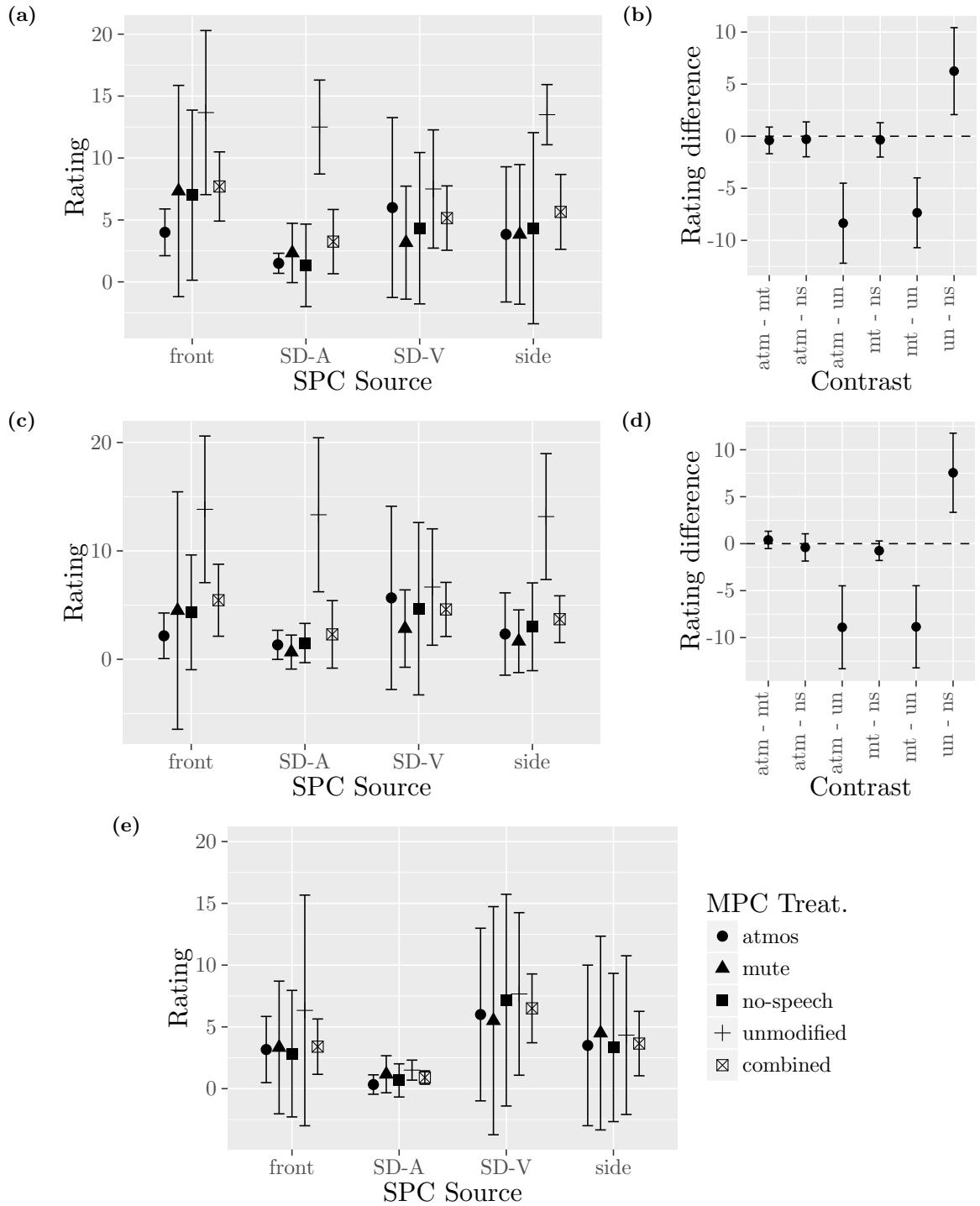
with the exception of the front and side conditions (SD-A:SD-V:  $t(19) = 8.472, p < .001$ ; SD-A:front:  $t(19) = 2.978, p = .020$ ; SD-A:side:  $t(19) = 3.018, p = .013$ ; SD-V:front:  $t(19) = 5.983, p < .001$ ; SD-V:side:  $t(19) = 4.088, p = .001$ ).

The disruption by the MPC to the SPC (Figure 7.10a) was significantly affected by the MPC treatment ( $F(3, 10.711) = 19.267, p < .001$ ). The SPC source and the interaction between the two factors, however, display no significant effects (SPC source  $F(3, 10.171) = 2.127, p = .159$ ), SPC source $\times$ MPC treatment  $F(9, 11.797) = 1.373, p = .300$ ). Pairwise comparisons of the MPC treatments indicate that the unmodified condition was rated significantly higher than the atmosphere ( $t(19) = -6.394, p < .001$ ), mute ( $t(19) = -6.542, p < .001$ ) and no-speech treatments ( $t(19) = 4.407, p = .001$ ). No other differences reach significance (see Figure 7.10b).

The ratings of the disruption to the SPC by the auditory content of the MPC (Figures 7.10c and 7.10d) exhibit a significant main effect for the MPC treatment ( $F(3, 11.386) = 16.823, p < .001$ ), but there is no significant effect for the SPC source ( $F(3, 10.609) = 0.835, p = .503$ ). Though the SPC source $\times$ MPC treatment interaction approaches significance ( $F(9, 11.787) = 2.664, p = .059$ ), the p-value exceeds the  $\alpha$ -level of this study (.05) and so the null hypothesis cannot be rejected. Pairwise comparisons of the MPC treatments indicate that the ratings were significantly higher for the unmodified condition than for the atmosphere ( $t(19) = -4.904, p < .001$ ), mute ( $t(19) = -5.567, p < .001$ ) and no-speech ( $t(19) = 4.929, p < .001$ ) conditions.

Ratings of the disruption that the visual elements of the MPC caused to the experience of the SPC (Figure 7.10e) indicate a significant main effect for the SPC source ( $F(3, 8.758) = 4.408, p = .037$ ) but non-significant effects for MPC treatment ( $F(3, 8.245) = 0.954, p = .458$ ) and for the SPC source $\times$ MPC treatment interaction ( $F(9, 10.826) = 0.378, p = .922$ ). Pairwise comparisons of the SPC source conditions across MPC treatments reveal a significant increase in the ratings for visual disruption between the SD-A and SD-V conditions ( $t(19) = 4.263, p = .002$ ). All other pairwise differences do not reach significance.





**Figure 7.10:** Plots showing the ratings of disruption caused to the experience of the SPC by the: (a)(b) MPC; (c)(d) MPC audio; (e) MPC visuals. (a), (c) and (e) show 20% trimmed means for each MPC treatment in each of the SPC source groups with 95% CIs. (b) and (d) show 20% trimmed means of participant difference scores between the MPC treatments (Rom's corrected 95% CIs).

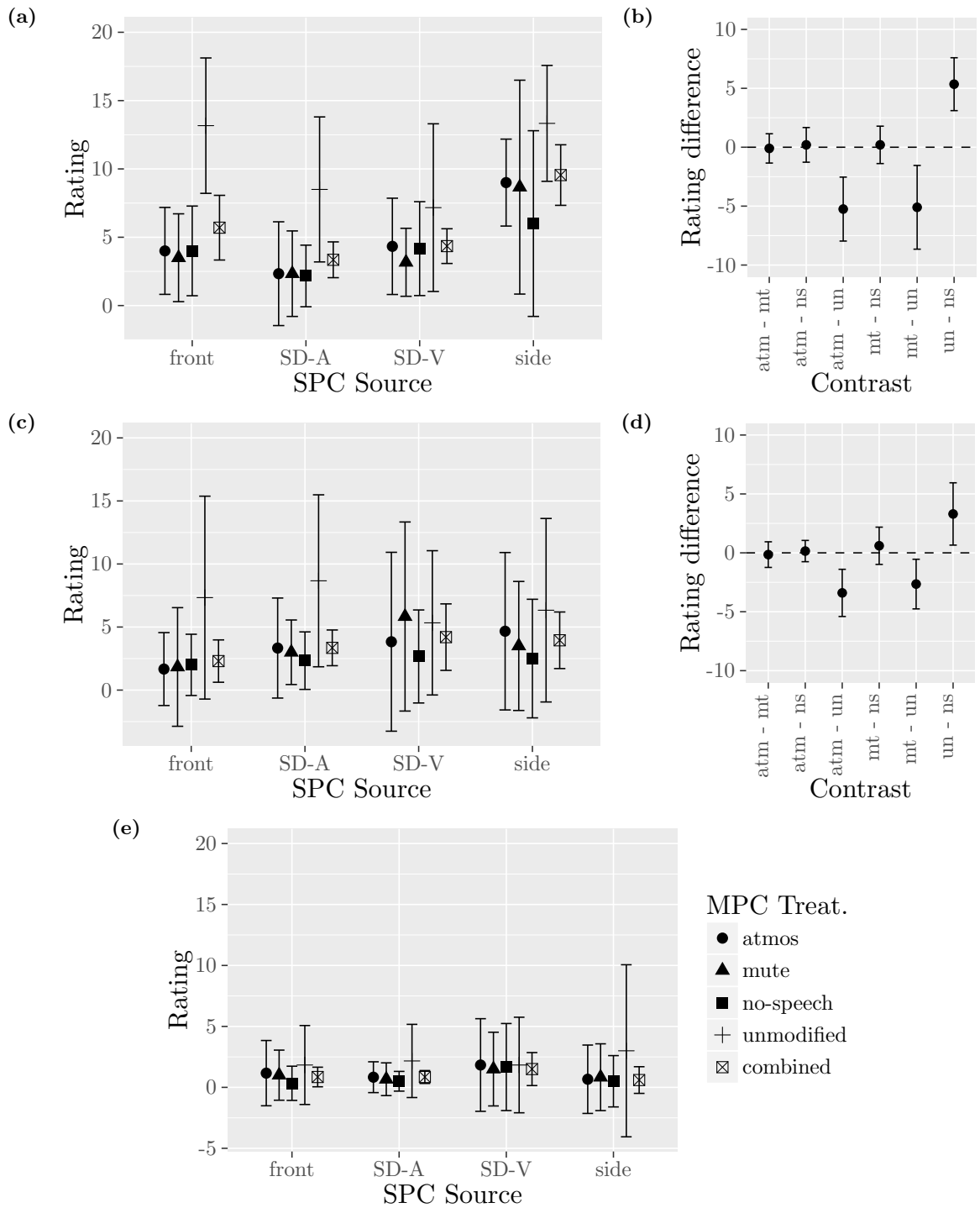
## Workload

Mental demand ratings (Figures 7.11a and 7.11b) exhibit significant main effects for both MPC treatment ( $F(3, 9.999) = 25.359, p < .001$ ) and the SPC source ( $F(3, 10.958) = 4.961, p = .020$ ). The effect of the SPC source  $\times$  MPC treatment interaction is not significant ( $F(9, 11.613) = 2.333, p = .089$ ). Pairwise comparisons of the difference between MPC treatment ratings pooled over all SPC source conditions show that the atmosphere ( $t(19) = -4.707, p = .001$ ), mute ( $t(19) = -3.949, p = .003$ ), and no-speech ( $t(19) = 6.561, p < .001$ ) conditions were rated as causing significantly lower mental demand than the unmodified condition. Pairwise comparisons of the SPC source ratings pooled over the MPC treatments indicate that when the SPC was presented from the side, it was rated as more mentally demanding than when it was presented from the secondary device, both as sound ( $t(19) = 5.201, p < .001$ ) and visually ( $t(19) = 4.395, p < .001$ ). The pairwise comparison of the front and side group scores also approaches significance ( $t(19) = 2.564, p = .055$ ).

Ratings of the temporal demand of the experiences (Figures 7.11c and 7.11d) were significantly affected by the MPC treatment ( $F(3, 14.110) = 7.976, p = .002$ ) but not by the SPC source ( $F(3, 10.980) = 0.169, p = .915$ ) or by the interaction between the two main effects ( $F(9, 12.035) = 0.784, p = .637$ ). Pairwise comparisons reveal that on average participants rated the temporal demand higher for the unmodified condition than for the atmosphere ( $t(19) = -4.131, p = .003$ ), mute ( $t(19) = -3.460, p = .012$ ) and no-speech ( $t(19) = 3.434, p = .011$ ) conditions.

Analysis of the ratings of physical demand (Figure 7.11e) indicates that neither of the main effects nor their interaction are significant (SPC source:  $F(3, 10.376) = 0.095, p = .961$ , MPC treatment:  $F(3, 12.265) = 2.257, p = .133$ , SPC source  $\times$  MPC treatment interaction  $F(9, 11.548) = 0.523, p = .831$ ).

Effort ratings (Figures 7.12a and 7.12b) were significantly affected by the MPC treatment conditions ( $F(3, 12.966) = 15.600, p < .001$ ). The figure shows that effects of the SPC source ( $F(3, 11.027) = 1.024, p = .419$ ) and the SPC source  $\times$  MPC treatment interaction ( $F(9, 11.801) = .853, p = .587$ ) are non-significant. Pairwise comparisons of the differences between participants' ratings indicate that participants rated the unmodified condition as requiring significantly more effort than the atmosphere ( $t(19) = -4.391, p = .001$ ), muted ( $t(19) = -4.184, p = .002$ ) and no-speech ( $t(19) = 5.126, p < .001$ ) conditions. No other comparisons reach significance.



**Figure 7.11:** Plots showing the ratings of (a)(b) mental demand; (c)(d) temporal demand and (e) physical demand. (a),(c) and (e) show 20% trimmed means for each MPC treatment in each of the SPC source groups with 95% CIs. (b) and (d) show 20% trimmed means of participant difference scores between the MPC treatments (Rom's corrected 95% CIs)

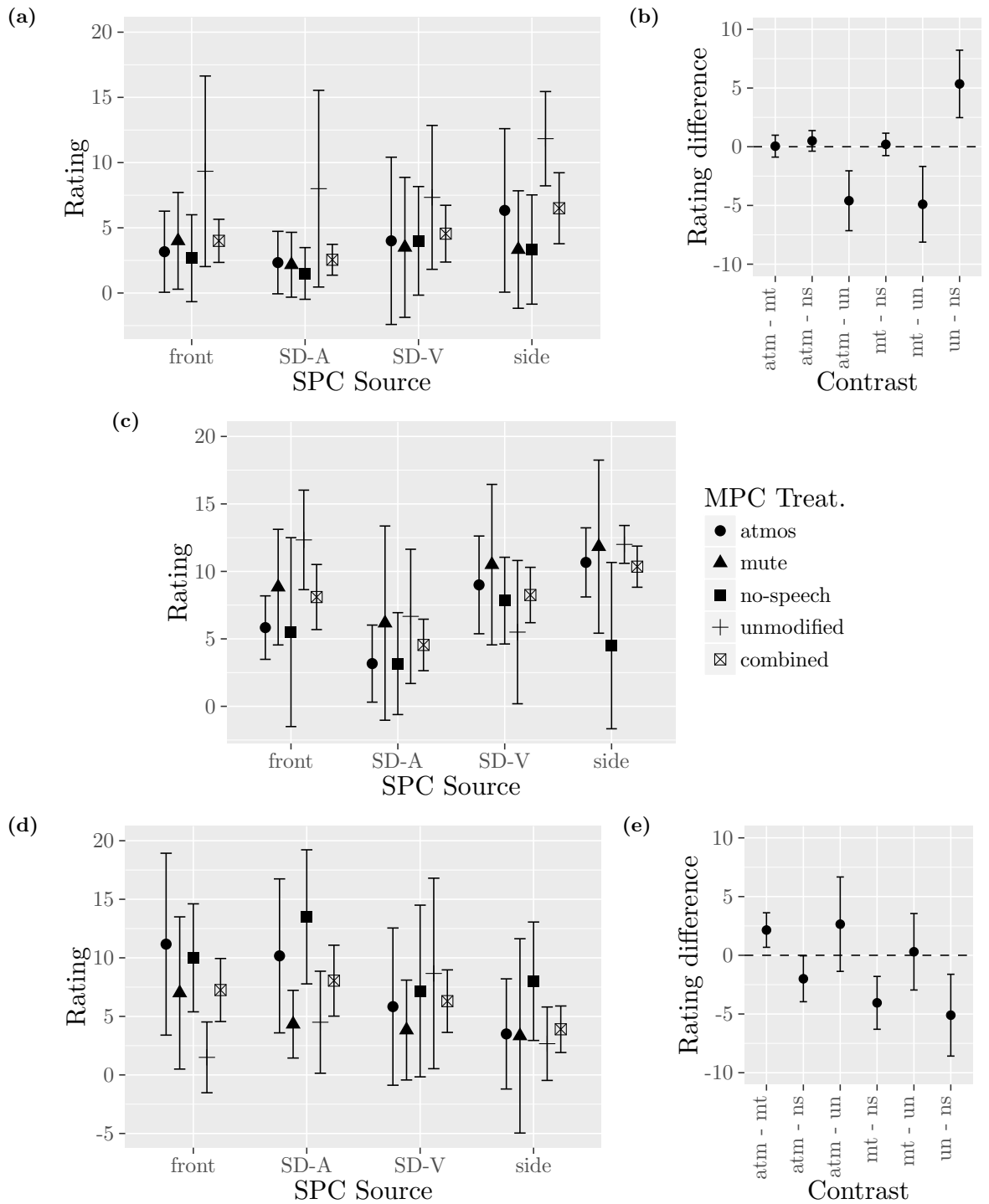
Annoyance ratings (Figure 7.12c) exhibit a significant main effect for the MPC treatment ( $F(3, 14.248) = 13.349, p < .001$ ), while the effect of SPC source approaches, but does not reach, significance ( $F(3, 10.996) = 3.379, p = .058$ ). The SPC source $\times$ MPC treatment interaction, however, is significant ( $F(9, 11.931) = 5.545, p = .004$ ). To explore the nature of the interaction, simple effects analysis was performed across the MPC treatment conditions for each SPC source with a one-way repeated measure ANOVA on the 20% trimmed means with a generalisation of the Huyn-Feldt correction for sphericity violations (Wilcox, 2012, p. 80).  $p$ -values were multiplied by the number of comparisons to correct for family-wise error. Results indicate that the MPC treatments did not result in significant differences in the ratings for the SD-A ( $F(1.514, 7.570) = 2.619, p = .568$ ) or the SD-V ( $F(2.022, 10.108) = 3.096, p = .356$ ) groups. Significant effects, however, are observed within the front ( $F(2.468, 12.340) = 6.738, p = .032$ ) and side ( $F(3, 15) = 6.060, p = .026$ ) groups. Visual inspection suggests that the unmodified conditions were rated as more annoying than the other conditions in the front condition (Figure 7.12c). In the left group, however, the no-speech condition was rated as less annoying than all other methods which appear to have been rated similarly to each other. Significant main effects were further investigated with pairwise comparisons of the differences in participants' ratings using the Hochberg correction for multiple comparisons to maintain the alpha value ( $\alpha = .0125$ ) for each simple effect. These comparisons revealed that none of the differences reach significance.

## Preference

The preference ratings, shown in Figures 7.12d and 7.12e, reveal a significant main effect for MPC treatment ( $F(3, 14.326) = 9.864, p = .001$ ), while the effects of SPC source and the SPC source $\times$ MPC treatment interaction are not significant ( $F(3, 10.912) = 2.100, p = .159$ ,  $F(9, 11.950) = 11.660, p = .204$ ). Pairwise comparisons of the differences in participants' rating scores indicate that the atmosphere condition was rated significantly higher than the muted condition ( $t(19) = 3.063, p = .024$ ). The no-speech condition was rated higher than the atmosphere ( $t(19) = -2.680, p = .044$ ), muted ( $t(19) = -4.715, p = .001$ ) and the unmodified ( $t(19) = -3.834, p = .005$ ) conditions.

### 7.4.6 Qualitative Results

A thematic analysis was performed on the free-text comments made by the participants following each clip. As the comments were optional, not all participants completed them for



**Figure 7.12:** Plots showing the ratings of (a)(b) effort, (c) annoyance and (d)(e) preference. (a), (c) and (d) show 20% trimmed means for each MPC treatment in each of the SPC source groups with 95% CIs. (b) and (e) shows 20% trimmed means of participant difference scores between the MPC treatments (Rom's corrected 95% CIs)

all clips. As balance is not such a concern within qualitative analysis, the two participants who were excluded from the statistical analysis due to procedural issues (discussed in Section 7.4.1) were partially included in the qualitative analysis. While comments from unaffected clips were included, those from the affected clips were not. In practice this impacted only one comment, as one of the participants refrained from commenting. While the comment made no reference to the fault that occurred, the possibility cannot be ruled out that this experience affected the response. The other participant who had been excluded due to being assigned the same order and condition as another participant was also included in the qualitative analysis.

The thematic analysis followed the stages outlined by Braun & Clarke (2006), comprising familiarisation, initial coding, identifying themes, reviewing themes, and writing up. The familiarisation took the form of reading the comments from the participants and making initial notes about the content. During the initial coding, each of the comments was reviewed and split into extracts (retaining the context of the comment), and coded to allow similar comments to be combined. This process was performed across all treatment groups to identify common patterns that emerged within the experience as a whole. Codes were then clustered based on similarities. This was initially performed by printing out the codes and arranging them to form a reconfigurable mind map. This was, however, superseded by a digital representation as the complexity increased and re-organisations proved increasingly difficult. The themes that this identified were then reviewed and refined against the original data corpus. After beginning the writing process, it became clear that some further refinements to the model were required due the degree of overlap between themes and also because one of the themes was over-dispersed. The final model was constructed and once more reviewed against the original data corpus to ensure that the themes captured the information present. This process resulted in eight final themes:

- **Disruption**—discussion of disruption, distraction or interference between the SPC presentation and the MPC
- **Missing out**—discussion of missing elements from one or both of the informational streams
- **Choice**—decisions about the stream(s) that were attended to and how easy participants found it to choose or attend to a source
- **Integration**—how effectively the SPC and MPC combined to provide the impression

of one experience

- **SPC Information**—opinions on the information conveyed by the SPC
- **Preference**—comments about preferences between clips or treatments
- **Suggestions**—suggestions or comments about features that were not included within the prototype
- **Miscellaneous**—comments which, though interesting, were not determined to correlate with other themes or to be strong enough to form a separate theme

The contents of each of these themes is discussed next and related to relevant extracts from the participants' comments. Due to the complex nature of the study, participants are associated with four individual comments, one for each MPC condition. The SPC group and MPC treatment are therefore provided along with the participant number following each quote, unless it is specified in the accompanying text. Clear typographical errors (incorrect spellings, case, or double spaces) have been corrected for the sake of readability, any other changes to quotes are indicated through the use of square brackets.

### Disruption

Within the theme of disruption, some participants discussed mutual disruption, where both SPC and MPC disrupted each other. Some comments from participants in auditory presentation groups suggested mutual disruption had occurred due to an acoustic interference between the two streams. Participants commented that the “[v]oices clashed far too much” [P14, *side, norm*] and that it “was difficult to hear both the secondary and main content” [P2, *front, unmodified*]. Others spoke of the distracting nature of the other stream (e.g., “The speech from the main content distracts me away from the secondary content and vice-versa” [P16, *SD-A, unmodified*]). It is notable that all of the comments within this sub-theme were made by participants talking about unmodified MPC conditions. Some participants referred only to one-way effects where one stream interfered with the other (e.g., “Voices from the main content made it difficult to listen to the secondary content” [P18, *SD-A, unmodified*] and “the speech from the second content interfered with the main program” [P10, *side, unmodified*]). A few participants commented that when the SPC and MPC speech did not overlap the disruptive effects were reduced (e.g., “When the content was presented during the non speaking portions of the clip, the secondary audio content was clear” [P9, *side, unmodified*]).

Conversely, one participant counter-intuitively claimed that they “found the second content to be much easier to absorb when multiple speech sources were interfering with each other” [P10, *side, atmos*]. It is suspected that this was not what the participant meant as they went on to say that “[t]he use of the phone and side speaker alleviated this issue somewhat”. Interestingly, one participant’s comment suggests that auditory SPC being displayed from the front location disrupted their experience of the visual MPC due to difficulty synchronising the two streams: “[t]he sound content was difficult to sync with the visual content and in some ways obscured it, and that is when I felt annoyed” [P6, *front, mute*]. Disruption from the MPC voices, was not solely limited to auditory presentations of SPC, however, as participants from the SD-V group reported that they “[r]eally felt the interference of the voices in the show with what I was trying to read on the phone” [P27, *unmodified*] and indicated that distraction caused difficulties, saying, “[f]ound I was switching between the main program and the secondary content, so didn[']t get much of either” [P24b, *SD-V, unmodified*].

Those who had experienced the modified MPC soundtracks also referred in their comments to disruptive effects (e.g., “felt like the main content was disrupted by losing the dialogue of the programme” [P14, *side, atmos*]). One participant highlighted the disruptive nature of being able to see a person talking in the visual MPC but being unable to hear them due to the MPC soundtrack manipulation, saying “[t]he soundtrack dropping in the main visual presentation felt very distracting. Especially when it was clear people were talking.” [P19b, *SD-A, mute*]. Interestingly, there was no reference to the continuation of music as being disruptive to the SPC, despite its potential for increasing distraction and reducing the SNR of the SPC. In fact, one participant explicitly indicated that it did not cause disruption to the SPC, saying that “[i]ncluding the music [...] didn’t make the secondary content any harder to follow” [P18, *SD-A, no-speech*]. The continuation of the atmospheric sounds in the atmosphere condition, however, was considered disruptive. One participant commented that “having the music drop out and only retain the background noise is somewhat jarring” [P8, *side, atmos*].

Furthermore, one participant from the SD-V group also noted that there was a delay after they finished reading but before the MPC modification was removed that “felt too long and a little distracting” [P29, *SD-V, no-speech*].



### Missing out

Feelings of missing out were a common theme in comments across the treatment conditions. Disruption can cause information to be missed and so there is an intrinsic link between these themes. The missing out theme, however, also encompasses comments that do not appear to be directly linked to disruption. It is considered separately here for this reason.

As for the comments on disruption, some participants described cases where information was lost from both streams (e.g., “I struggled to take in information from either the main or secondary content” [P5, *front, unmodified*]). These comments have a strong link to mutual disruption, which led to many common extracts existing between the themes. In some cases, interference between sources was identified as being the cause of missing information (e.g., “I could not experience either content due to the disruption of them both on each other” [P21, *SD-A, unmodified*]). Interestingly, one participant talked about it being “hard to remember what either audio was saying” [P1, *front, unmodified*]. This wording would seem to suggest that the problem of missing out on the information was experienced when trying to recall the information, rather than when initially attending to the sources.

Many participants expressed concerns about missing content from the MPC during the experiences due to the manipulation of the soundtrack. This issue clearly affected participants across the SPC presentation conditions, with comments from all of the four SPC source groups. Participants commented that they felt they were “missing out on information from the clip itself as some of the sound was muted to avoid distraction.” [P26, *SD-A, atmos*], and that they “did not like the way the sound cut out when I had secondary text on my phone in land of the jaguar, I thought I was missing out” [P24b, *SD-V, mute*]. Participants also expressed feeling that they wanted to hear elements of the MPC but were unable to, indicating that they would rather have heard the MPC (e.g., “I wanted to hear what was going on in the main content more than the secondary” [P18, *side, atmos*] and “I would have rather got back to the main program a bit quicker” [P14b, *side, atmos*]).

Many participants specifically referred to concerns about the loss of the information from the MPC speech, saying that they “felt a little like I might be missing out when voices were cut out” [P18, *SD-A, no-speech*] or “missed what was being said” [P15, *side, no-speech*]. It is likely that being able to see on-screen talkers in the MPC played a large role in these comments, as in the no-speech condition this would have been the only cue to alert the viewer that voices were removed. Some explicitly referred to seeing people talking but being unable to hear them as being a cause of annoyance (e.g., “when people are on screen talking, it is

annoying not being able to hear what they are saying” [P2, *front, no-speech*]). The presence of muted on-screen talkers obviously brought about feelings of missing out that affected the participants feelings towards the MPC modification (e.g., “This form of presenting the secondary content only works in certain contexts - while it works very well for landscape shots, etc., it is quite jarring when you can see we are supposed to know what someone in the main program is saying, for example in a talking head interview” [P8, *side, no-speech*]). A couple of participants raised the concern that the MPC modification would cause the user to miss narrative information (e.g., “there is the danger that the narrative of the program can be lost to the viewer since the main soundtrack ducks out” [P7, *front, atmos*]). There was, however, some indication that the inclusion of music helped to reduce the feelings of missing out, with participants commenting that “It still felt a little like I might be missing out when voices were cut out, but the music helped minimise this effect” [P18, *SD-A, no-speech*], and “I felt as if I was not missing much from the primary media as I was not losing the atmospheric music” [P1, *front, no-speech*]).

Feelings of missing out were dependent on the content of the MPC. One participant reported that the visual MPC reassured them that important narrative information was not interrupted, saying: “it did not feel like I was missing any vital information from the main program, as the researchers were simply performing some manual labour and so their progress and the narrative of the documentary was not interrupted” [P10, *side, no-speech*]). Several participants also referred to feeling as though they were missing out on the MPC at a point where the team discovers that they have captured footage of a jaguar and this footage is displayed. Participants within the SD-V group were clearly affected by this (e.g., “[f]elt like I missed an exciting bit with the jaguar as I was reading the secondary content” [P24b, *atmos*]). Comments about this section were not restricted to participants in the SD-V condition. One participant from the *side* group referred to the MPC modification as leading them to feel as though information had been missed: “the secondary content cut out the speech from the main program at an interesting point when the jaguar was discovered, I therefore felt I was missing exciting information.” [P10, *mute*]). One participant spoke about feelings of missing out in the unmodified condition being dependent on the type of spoken content that occurred at the same time as the SPC, with the narration considered as being more important and therefore a greater loss. They commented that “when it was just the people talking because I could distinguish between the secondary content and the video quite easily and it was fine when there wasn’t much dialogue, but the secondary content annoyed me when the narrator was speaking because I was missing out on important information about what was happening

to listen to background stuff” [P15, *side*]. The comment appears also to suggest that the dialogue suffered less interference from the presence of the SPC than the narration. The comment is considered also to fall within the theme of disruption.

### Choice

Most of the comments within the choice theme refer to a participant attempting to focus on, attend to, or ignore one of the two streams of content. Some participants described difficulties in attending to one of the streams. These comments largely also fall into the theme of disruption, as participants described the other stream that affected their ability to focus on the chosen stream as a distraction and interference (e.g., “It was difficult to hear both the secondary and main content, therefore it was impossible to follow either” [P2, *front, unmodified*]). A couple of participants described switching between the MPC and SPC and it leading them to miss information (e.g., “Found I was switching between the main program and the secondary content, so didn[']t get much of either” [P24b, *SD-V, unmodified*]) and “I felt like I was missing out on key information as I was flicking between listening to each source.” [P14, *side, unmodified*]). It is notable that these two participants were from auditory and visual SPC representations, suggesting that there were similar elements of attentional conflict present in both conditions.

One participant described switching between focusing on the main and secondary content as being “difficult and unenjoyable” [P14, *side, unmodified*]. Conversely, one participant indicated that they were able to selectively attend to, and switch between sources. They said that the unmodified condition allowed them to “choose which one I wanted to “tune into”, so I felt like I had more control of the experience - if the secondary content turned out to be boring I could concentrate on the main show” [P13, *side, unmodified*]. Another participant, however, apparently struggled with choosing which stream to attend to “It was hard t[o] tell which I was ‘supposed’ to be listening to” [P18, *SD-A, unmodified*].

One report suggested that in the SD-A condition, the MPC was more easily attended to than the SPC, as one of the participants commented that they “[f]elt it was easier to concentrate on the main content, information seemed more relevant but harder to concentrate on.” [P22, *SD-A, unmodified*]. This effect may have been due to the reduced fidelity of the secondary-device’s speaker, or a lower sound intensity of this source at the ear due to the comparatively low power of the phones speaker and/or the orientation of the secondary device. Many of the comments from this theme were made by participants who had experienced

auditory representations of SPC being presented alongside unmodified MPC. While very few spoke about choosing between the speech sources explicitly, the fact that the comments are largely restricted to this condition suggests that the participants were generally referring to choosing between speech sources. It is also notable that the participants felt they had to choose a source in this scenario. One participant commented that it was “impossible to follow the main program and secondary content at the same time” [P8, *side, unmodified*], suggesting that dividing attentional resources between both streams at the same time was either impossible or very difficult.

In the SD-V treatment group, one participant referred to the SPC presentation that occurred at the same time as the jaguar footage (further discussed in the ‘missing out’ theme) and chose to attend to the main screen instead: “secondary content occurred at the same time as the interesting footage of a jaguar on the main content, I didn’t want to read the secondary content and instead focussed on the main.” [P29, *SD-V, atmos*]. Interestingly, there was also some indication of participants choosing to ignore auditory SPC in modified MPC treatments, as one participant talking about the secondary content said “I can block it out as I see fit” [P6, *front, no-speech*].

### SPC Information

The information conveyed as SPC was not to everyone’s taste. Some participants did not consider parts of the SPC interesting (e.g., “the subject of the second content is not always of interest to me” [P10, *side, unmodified*]). Others felt that the SPC was irrelevant or unnecessary (e.g., “Some of the information felt irrelevant which made it more annoying” [P22, *SD-A*]). One participant went so far as to say that “[i]f I wasn’t in an experiment I probably wouldn’t bother reading the secondary content” [P28, *SD-A, atmos*]. Some users, notably within auditory conditions, felt that the SPC segments “went on too long” [P8, *side, mute*] and that they “would prefer shorter bits of information” [P12, *side, no-speech*]. While these negative comments were spread across treatment groups and clips, two participants specifically highlighted their dislike of the biographical information (i.e., “the secondary content about the presenter’s background wasn’t of interest” [P23, *SD-A, mute*], and “if the secondary content is always a recital of someone[?]s CV I probably wouldn’t be inclined to switch it on” [P10, *side, no-speech*]).

One of the participants recognised the potential to personalise these experiences by providing a choice of topics (“I could see though that multiple types of secondary content could be

attributed to the same program” [P10, *side, no-speech*]) and allowing users to select them (“but I could presumably decide the subject matter I want to be notified about” [P10, *side, unmodified*]). Another suggested that the information should be split differently between the MPC and SPC streams, saying “I prefer the biographical information to come from the individuals or narrator with the more technical information about cameras and the rainforest to be provided in the secondary content” [P5, *front, mute*].

Some participants spoke positively of the SPC information. In contrast to the negative comments, they found the SPC information interesting (e.g., “Interesting to get background info to set the scene” [P31, *SD-V, atmos*]). Others said “the secondary content was useful” [P6, *front, mute*], that it “complemented the main programme” [P5, *front, atmos*] and that “it felt like it was adding to my experience of the clip” [P26, *SD-V, atmos*]. One participant also demonstrated that the negative feelings towards biographical SPC were not universal, stating “Nice to know more about the forest and the camera-women” [P31, *SD-V, unmodified*]. None of the positive comments about SPC information were from the third clip, while it was commonly referred to amongst the negative comments. It is possible that the SPC information presented in this clip was generally less liked than in the others. As comments were voluntary, however, it is not possible to be certain that this was the case or whether others with positive, or non-negative views on the SPC in these clips simply chose not to comment on this element of their experience.

### Integration

Some participants described experiences of separation within the comments relating to integration, in which the experience of the SPC felt isolated or disconnected from the MPC (e.g., “Looking at the phone, sometimes I almost forgot I was watching TV. It felt like I’d completely switched to doing something else” [P27, *SD-V, mute*], and “I found that it removed myself from the experience of the program” [P14, *side, mute*]). Almost all of the comments in which participants referred to separation occurred within the mute condition. One participant likened the muted condition to a serial experience, saying “Having no soundtrack/voices from the main content at all felt like I was experiencing one content then the other, not both at the same time” [P29, *SD-V, mute*]. One comment following an atmosphere treatment clip, highlighted the removal of music as an important factor: “Taking away the music when the extra information is given takes away from the atmosphere of the programme” [P12, *side, atmos*].

Feelings of separation are contrasted with descriptions of integration described in other parts of the data, where participants spoke about it feeling “like a more cohesive experience rather than two separate pieces of media” [P29, *SD-V, unmodified*], or “more integrated” [P22, *SD-A, atmos*]. Some of the participants highlighted the continuation of some MPC audio as being the reason for this impression of integration (e.g., “The secondary content still felt part of the main programme because of the continuing atmospheric sounds.” [P23, *SD-A, no-speech*]). No comments regarding feelings of integration originated from muted trials. This adds to the impression that the continuation of ambient sound during the SPC presentation was an important factor. The one comment concerning an unmodified condition, originated from a participant in the *SD-V* group, where one would expect there to have been less disruption between the SPC and MPC voices.

The muting of on-screen talkers proved to be a negative factor within the theme of integration. One participant suggested that the presence of a muted talker reduced the feeling of integration saying “it felt more like part of the programme with the continuation of the sound - except when it was clear that someone was speaking on screen” [P22, *SD-A, no-speech*].

## Preference

The preference theme predominantly consisted of participants making comparisons between the various treatments they had experienced and commenting on which they preferred. Due to the sparseness of this data and the presence of the rating scale that explored participants’ preferences as part of the quantitative analysis, these comments were not subjected to in-depth analysis. Furthermore, many comments went on to describe reasons for preference in terms of disruption, integration or missing out and therefore are duplicated across themes. There were, however, a few comments within this theme that were considered to add insight. Of all of the comments there were only two that presented an opinion on the source of the SPC. One participant from the *SD-V* group referred to a preference towards the unmodified condition over the no-speech and atmosphere conditions: “because so many times people use their phones while watching a programme, this is much more comfortable for us to do, and while it may seem slightly more disruptive, it was far less irritating” [P26]. The other was from a participant in the front group, who commented: “The notion of presenting the secondary content at the same level and spatial direction as the primary soundtrack is ridiculous” [P7, *unmodified*].

## Suggestions

Several participants suggested alternate features that were not provided as part of the prototype experience. The use of subtitles within the experience was suggested by a few participants. One recommended the use of subtitles as an alternative to auditory presentation, commenting “we are more used to subtitles and I would have welcomed the secondary content in that way” [P6, *front, mute*]. This comment may also be seen as the participant expressing a preference towards a textual representation of SPC over auditory presentations. It is interesting, however, that they refer to the use of subtitles, despite having experienced the SD-V condition as part of the training session. This therefore suggests a preference towards a text-overlay on the main screen, rather than the use of a second screen. Alternatively, there was some suggestion of using subtitles to compensate for the removal of the voice in the MPC modified conditions (e.g., “It would have been better if when the secondary content was playing, the dialogue from the show was shown in subtitles on the screen” [P15, *SD-A, no-speech*]).

There was a suggestion of interleaving the SPC into pre-existing dialogue gaps: “It would have been better to be able to have listened to the person speaking on the main program and have the secondary content only when a person is not speaking and music is playing whilst a clip is being shown” [P11, *side, no-speech*]. Whilst another participant suggested that it would have been better if specific elements of the soundtrack were attenuated rather than being completely removed saying: “A drop of even 6-10dB in the primary soundtrack would be enough to reduce the irritation caused significantly” [P7, *front, unmodified*], and “merely reducing the level of the soundtrack would be sufficient” [P7, *front, mute*].

One participant suggested an alternative context of use for the SPC. After pointing out they would not use the display “if it distracted from the main program especially if important narrative conversations were taking place” [P10, *side, atmos*], they pointed out that “[i]t may however be interesting to watch a program I really like a second time with additional secondary content that locks out the main narrative”. This suggestion appears to refer to an experience similar to that provided by the ‘director’s commentaries’, which are commonly included as bonus features on DVDs and Blu-Rays. The narrative of the MPC is less important in these contexts, as it is assumed that the user has previously seen the main programme.

## Miscellaneous

Several participants from different conditions expressed negative feelings about a talker in the MPC being interrupted by SPC during the third clip (e.g., “When the audio cut out while the woman was speaking I did not like it” [P24b, *SD-A, no-speech*]). It seems likely that the instance the participants were referring to is where SPC was timed to start during a short pause in some on-screen dialogue. It is clear, however, that participants still felt that the talker was interrupted. It is interesting to note that all of the comments associated with this were attributed to MPC treatments in which the dialogue had been removed. This suggests that users who experienced the clip as unmodified did not feel that the MPC was interrupted as much or, alternatively, that other factors within their experience were more significant.

There was another cluster of comments, which spoke about effects of continuing some of the MPC audio elements throughout the SPC display. Several of these discussed the enhancement to the experience from the continuation of the music (e.g., “With musical accompaniment, the secondary audio was much more pleasant than previously and was almost relaxing” [P1, *front, no-speech*]). The continuation was not received well by all, with one participant in the atmosphere condition saying “[o]nly retaining background noise from the main program seems pointless more than anything else” [P8, *side, atmos*]. One participant raised an interesting concern about the continuation of atmosphere alongside the music when the voices had been removed. They commented that keeping the atmosphere at its normal level “could perhaps be too revealing about the soundtrack design and end up becoming annoying in itself when the main dialogue resumes (i.e. the illusion is broken!)” [P7, *front, no-speech*]. This may be taken to suggest that the atmosphere, in addition to the dialogue, should be removed to maintain the impression that all of the sounds that are supposed to be present within the world portrayed in the programme (diegetic sounds) are a single entity. Unfortunately, this condition was not included as part of this experimental design and so it is not possible to comment on how this variation would have changed the experience.

There was also criticism of removing all of the sound, with one participant remarking on how the modification “made the main content feel secondary to the secondary content” [P13, *side, mute*]. Two participants apparently became disengaged with the experience during the muted conditions (i.e., “It felt a bit awkward to just go silent and listen to the secondary content, It made it easy to switch off” [P21, *SD-A, mute*] and “I lost all interest in what the secondary content was telling me and wanted to get back to hearing the original voices/music/etc from the main program” [P14, *side, mute*]).



The delay between finishing reading VSPC and the MPC treatment being removed was highlighted by one participant under the theme of disruption. A couple of participants also referred to the irritation it caused (e.g., “I did find it irritating when I’d finished reading but the voices and music were still muted” [P27, *SD-V, atmos*]). This issue is primarily due to the lack of interactivity within the prototype system that was used for testing. It seems likely that by including the ability to return to the main programme, and removing any soundtrack modification, these issues would be avoided. A few participants spoke about issues of control with the experience. While one participant described how they “felt out of control” [P13, *side, atmos*], others acknowledged the benefit that interaction would give to the experience. For example, one participant talked about the inconvenient timing of some SPC and went on to say “[t]herefore the interactive element would make this experience better” [P10, *side, mute*], while another commented “If it was optional would have no problem with this” [P23, *SD-A, atmos*]. A couple of participants compared the experience to pausing. One commented that “[i]f this was how the secondary content was delivered, in real life I would just pause the programme if possible” [P23, *SD-A, unmodified*]. The other participant commented following a muted trial that “you may as well just pause the programme and play the additional content” [P12, *side*].

The experience may have taken some time to get used to. One participant commented in the third of the four experimental trials that they were “beginning to get more used to the secondary content, beginning to expect it and factor it in” [P6, *front, no-speech*]. This comment would appear to suggest that, despite the familiarisation session, some of participants were still getting used to the display. This may have been partially due to the lack of experience with the MPC treatments.

One participant commented that one of the pieces of SPC “seemed unnecessary- the lady could have been saying this” [P23, *SD-A, mute*]. This indicates that there was some doubt about the amount of redundancy between the information in the SPC and MPC streams. A couple of participants also referred to using third party web services as an alternative to modifying the MPC for the ASPC. One remarked “if I wanted to know everything about Steve thingy, I would google him and read it rather than missing part of the show to hear the contents of his wikipedia page” [P15, *side, mute*], while the other commented “I think this unnecessary when it mutes the primary media’s audio because such information is somewhat readily available on wikipedia, which with a smartphone I could access anyway” [P1, *front, mute*]. These comments could also be considered to indicate a preference towards a textual

representation of SPC.

There was one particularly anomalous comment in which a participant remarked about the presence of “three levels of content” [P6, *front, unmodified*]. It is not entirely clear what levels of content they were referring to here. Nothing was noticed when running the session, or from revisiting the video from of the session that could account for the perception of three levels of content.

One participant from the SD-A condition remarked “I often found myself looking at the phone screen even though it was blank.” [P18, *atmos*]. It is unclear from this statement whether the participant was looking at the secondary device because it was the source of the SPC, or because they expected there to be some content appearing on the device’s screen. The secondary device’s sound quality was criticised in one comment: “the sound quality of the phone speakers diminished the experience” [19b, *SD-A, no-speech*]. It is interesting that the quality of the secondary device sound was not raised by more participants. It may, however, have been a contributing factor to the reports of the SPC being harder to attend to (discussed in the choice theme).

### 7.4.7 Discussion

#### Effects of MPC treatment

Out of the MPC treatments, the unmodified condition appears to have been the least successful. For all of the ratings of disruption that were not specifically referring to the visual aspects of the MPC, the unmodified condition was rated as being significantly more disruptive than the no-speech condition and, with the exception of the disruption from the SPC on the MPC audio, the atmosphere treatment. It seems likely that this difference is largely due to the effects of the speech present in the MPC for the unmodified treatment. As an indication of the severity of the disruption during the unmodified treatment, for the ratings of disruption caused by the SPC to the MPC and MPC audio, participants’ ratings did not differ significantly to those for the muting of the main content. The unmodified condition was also rated as significantly worse than all of the other conditions in terms of mental demand, temporal demand and effort. The presence of MPC dialogue will have meant that participants will have had to attempt to ignore one stream in order to process the other, which may account for the increased effort and mental demand scores. While all of the participants experienced the clips at the same rate, the presence of the dialogue may have

meant participants felt that they had to process it more quickly so that they could return to attending to the MPC speech.

The comparatively poor performance of the unmodified condition is somewhat surprising in light of the work of Guerreiro & Gonçalves (2016) on the Text-to-Speeches system, which indicated that sighted participants were comfortable with scenarios involving the presentation of two streams of speech. There are a number of differences which may explain this apparent difference. Firstly, Guerreiro & Gonçalves (2016) maximally separated the competing streams in the horizontal plane so that sources were presented on opposite sides of the head. This is likely to have meant that attending to only the ipsilateral ear led to more a favourable TMR than in the present experiment, where there was less spatial separation. The scenario of use was also very different in the two studies. While Guerreiro & Gonçalves (2016) asked participants to try to attend to a specified stream of speech, in the present study participants were not explicitly instructed to attend to either source. This may have made the present study more difficult, as participants had to decide which stream to attend to. Also, while in Text-to-Speeches both sources comprised only speech, in the present study the MPC also included music, atmospheric sounds and effects. These components may have made the present task more difficult by contributing additional energetic masking and adding a potential further distraction. The presence of visual information corresponding to one of the streams in the current experiment is another important distinction. Visual information is known to affect the perception of audio (as discussed in Sections 3.2.6 and 3.3.4). It is possible that having visual information corresponding to the MPC made it more difficult to switch between streams than if only auditory information had been present, as in Guerreiro & Gonçalves's (2016) Text-to-Speeches system.

The muted condition was associated with more disruption by the SPC to the MPC and by the SPC to the MPC audio than both the no-speech and atmosphere treatments. These ratings suggest that leaving some of the programme audio reduced the degree of perceived disruption to the main programme despite the absence of the dialogue. Participants clearly felt as though these elements were adding to their experiences, and this is reflected in the qualitative comments under the theme of integration and the miscellaneous comments regarding the continuation of sound elements. The muted condition was not, however, found to significantly differ from ratings of disruption in the other direction. This suggests that participants did not generally consider the SPC to be significantly more disrupted by the presence of the atmosphere and music, or by the atmosphere alone, than complete silence.

Results from the preference scale indicate that the clear favourite was the no-speech treatment, which was rated higher than all other MPC treatments. While the preference for the no-speech over the unmodified treatment and the muted condition may be seen as a logical extension from the disruption and workload-style questions, no difference is present between the no-speech and atmosphere treatments for any other ratings. The qualitative analysis, however, sheds some light on the factors that influenced this preference. Within the miscellaneous theme, participants discussed the benefits of retaining music during the MPC in terms of engagement and enhancement to the SPC. Furthermore, there were a couple of comments under the missing out theme that referred to music reducing the feeling of missing out. Additionally, one participant commented that the removal of music in the atmosphere condition was disruptive. There were, however, no significant differences for any of the disruption measurements between the atmosphere and no-speech conditions. This suggests that the music in the MPC did not significantly impair the participants' ability to attend to the SPC. This assertion is supported by the comment from the qualitative analysis that identifies the music as not negatively impacting on the SPC.

A similar logical extension explains the difference between the muted and atmosphere conditions in the preference rating. The lack of a significant difference between the unmodified and atmosphere conditions, however, is more difficult to account for because of the consistent difference in disruption and workload-style ratings between the two. The lack of significance appears to be due to relatively large confidence intervals associated with the difference in preference ratings compared to the other difference scores. This suggests that different participants' opinions varied considerably on this. The reasons for this are unclear.

Despite the generally positive results from the quantitative analysis regarding the no-speech and atmosphere conditions, the qualitative analysis does highlight a couple of issues. Seeing on-screen characters talking but not being able to hear what they were saying clearly had a negative impact on the user experience. This may be because seeing someone speaking but being unable to hear anything is an inherently unnatural experience. Furthermore, the on-screen talker will have acted as a visual cue to the participants that they were missing information from the main programme. With the ideas of object-based broadcasting, it is interesting to ponder the effect that replacing the clips of the presenters talking with other shots without an on-screen talker would have had on participants' responses. This would become problematic, however, when considering a shared experience in which some users may not wish to access the SPC. In some circumstances, it may be possible still to present

the dialogue originally from an on-screen talker over different visuals (e.g., during a piece to camera). There are many instances, however, where such an approach would be problematic, such as during a conversation between two characters in a drama or if the talker is referring to something happening within the same shot. The use of this approach would require a great deal of caution.

The lack of a significant effect of MPC treatment for the disruption to, or by, the MPC visuals is likely to be due to the fact that the treatments themselves did not alter the ability to visually attend to the source.

### **Effects of SPC source**

The statistical analyses for SPC source was subject to more noise than the MPC treatment factor, due to it being a between-participant variable. This will have meant that random variation in the way that specific participants rated experiences (i.e., a bias towards ratings in a particular area of the scale) will have had a larger effect than for the within-participant variable. Furthermore, as participants were experiencing variations of the MPC treatment, they may have been more inclined to use more of the scale to express differences between these treatments rather than considering the absolute rating value. Several significant effects were nevertheless uncovered.

It would appear that the ‘side’ treatment was associated with several negative attributes by the participants. The ratings for disruption caused by the SPC on the MPC and the mental demand were significantly higher than for either of the handheld conditions. Additionally, the ratings for the side group can also be seen to be higher than the front group in both instances, though the difference was not significant in either case. The side condition was also found to be associated with the highest ratings for the disruption by the SPC to the MPC audio compared to all the other conditions, though despite a significant main effect, no pairwise comparisons reached significance. These results are particularly interesting, as one might expect a laterally positioned source to have been less disruptive than the front condition, particularly in the unmodified MPC treatment, due to the spatial-release from masking.

The finding that the ASPC presented from the secondary device’s was considered to be less disruptive to the MPC and less mentally-demanding than when presented from the side may be due to a number of factors. Firstly, the levels of the auditory SPC experienced at the ear in this condition are likely to have been considerably lower than those originating from the lateral

speaker. This level difference will have led to a reduced amount of energetic masking to the MPC audio compared to the other auditory SPC sources. It is likely to have been particularly important in the unmodified treatment, as both MPC and SPC streams contained speech. A reduced level may have also reduced the amount of informational masking to the MPC by the SPC, as the difference in level would allow participants to ignore the SPC stream more easily. Informational masking, however, would be expected to mainly affect the unmodified condition due to the presence of competing speech sources. Inspection of the results of the side and SD-A conditions indicates that this was not the case, as all MPC treatments show reductions in the SD-A condition with larger differences between trimmed means for the atmosphere and no-speech treatments. The lower presentation level of the SPC in the SD-A condition would be expected to cause participants to find it more difficult to attend to than for the fixed conditions. There is limited support for this from the qualitative analysis, with participants from the SD-A group discussing the disruption of the MPC voices on the SPC and the impression that it was easier to attend to the MPC than the SPC. This, however, is not evident in the ratings of disruption caused by the MPC audio to the SPC, for which no significant effect of SPC source is observed. The results suggest that the presentation did not greatly affect most participants' experiences of the SPC stream. Lower presentation levels of a target stream have previously been found either to have little effect or, in some cases, to be beneficial (Brungart, 2001; Brungart & Simpson, 2007; Ihlefeld & Shinn-Cunningham, 2008b). This may explain why the SD-A condition was not associated with higher ratings of disruption by MPC audio to the SPC compared to the other ASPC conditions, which were of similar levels to the MPC narration.

As the participants had control over the secondary device's location, this condition may have allowed them to position the phone so that it was in a more optimal position than the side location for all of the MPC treatments. The position in which the device was held was unlikely to correspond with the same location as the central speaker, meaning there will have been some spatial release from masking under the unmodified condition. Furthermore, participants had dynamic control over the location throughout each SPC presentation in the SD-A condition. This means that if they lost interest in the SPC they could re-position the source to minimise the disruption it caused to the MPC. Alternatively, if the SPC was of interest, they could reposition it into a more easily attended position. Such an approach would explain reduced mental demand, as participants were not struggling to attend to a chosen source and disruption to the MPC was reduced. From watching the videos of participants during the SD-A condition, it was difficult to ascertain in many cases whether movements

of the secondary device were due to unrelated shifts in posture. During the unmodified treatment, however, a few participants made notable changes in how they positioned the secondary device, lifting the phone up closer to the head (around chest height). This demonstrates that some of the participants made use of their ability to dynamically alter the devices location.

There are also a number of explanations for why the side location may have been perceived as particularly disruptive and demanding. Presenting the SPC from the side location may have been disliked because of the source's spatial separation from the locus of visual attention. Spence *et al.* (2000) found that it is more difficult to ignore a speech stream originating from the locus of visual attention in concurrent speech scenarios. This would suggest that in the unmodified condition that participants will have found it more difficult to ignore the MPC, if they chose to do so. Furthermore, in the modified conditions, participants may have found that the spatial separation led to the SPC feeling less well integrated with the MPC soundtrack, which therefore increased the perception of MPC being disrupted.

Alternatively, the greater disruption and mental demand could be due to sounds originating from the side being perceived as more salient and hence more attention grabbing than those from the front. Some researchers have found that sounds presented from behind generally lead to greater emotional responses (Tajadura-Jiménez *et al.*, 2010) which tend to be more negative than for sources positioned in front (Asutay & Västfjäll, 2015). These studies suggest that the results are due to the sounds originating from outside the visual field of view and the auditory system's role in orienting visual attention to assess them. The side position within the current study was 90° to the left of the participant. This location, would be classified as within a participant's field of view - generally considered to be 200° (DeValois & DeValois, 1990). At 90° from the median plane with the participant looking ahead, however, visual acuity would be low (Findlay & Gilchrist, 2003). If the negative emotional response associated with sounds originating from behind is due to the auditory system attempting to detect threats that would be missed by the visual system, it seems logical that this effect would also extend to extreme lateral locations. As both Tajadura-Jiménez *et al.* (2010) and Asutay & Västfjäll (2015) only tested locations on the median plane directly in front of and behind the listener, however, it is not possible to draw definite conclusions based on their findings.

The absence of SRM within the data may be attributable to several factors. Previously, some researchers have found moderate SRM when concurrent voices are qualitatively different

(Noble & Perrett, 2002), though this finding has not been universal (Allen *et al.*, 2008). It is possible that this was a factor in reducing the benefit of SRM in this study, as the voice of the SPC was deliberately chosen to be of opposite gender to the narrator in the MPC.

Additionally, participants were not exposed to two localised sources, but to two localised speech streams and a mix of music, atmosphere and effects that were spread around the 5.1 set-up. The energetic masking advantage of spatial separation, therefore, may have been reduced by the presence of non-speech audio from the other ipsilateral speakers.

Furthermore, there will have been some spatial difference between the front ASPC speaker and the centre channel of the surround sound. This separation, though small ( $\approx 6^\circ$ ), is still above the minimum audible angle in elevation of approximately  $4^\circ$  (Perrott & Saberi, 1990). SRM due to elevation differences close to the median plane have been observed with speech (Worley & Darwin, 2002; McAnally *et al.*, 2002; Martin *et al.*, 2012). These effects were, however, observed at larger elevation differences, the minimum being of  $19^\circ$  reported by Worley & Darwin (2002). As elevation-based SRM has been largely attributed to informational masking (Martin *et al.*, 2012), it is possible this small separation still had an effect. A release from masking may have reduced the difficulty of attending for the centre condition and so reduced the potential benefit from the side condition. Many of the MPC treatments involved the removal of the MPC dialogue and did not involve concurrent speech. It is therefore noteworthy that SRM would most likely be observed between the front and side SPC sources in the unmodified treatment.

One possible explanation for the lack of SRM could be insufficient statistical power for the interaction term. Inspection of the results, however, shows very little difference between these scores for all measures. Additionally, the methodology used within this study differs considerably from those which focus on the study of masking effects. Here, the focus was on the effects of treatments in a user's experience of clips as a whole, as opposed to studies on speech masking, which generally measure success at detecting or recognising individual words (e.g., Brungart, 2001; Ihlefeld & Shinn-Cunningham, 2008b; Martin *et al.*, 2012). Furthermore, the semantic contexts of the words used in these studies are often uninformative. Within this study, however, no specific consideration was given to this. It is therefore likely that when participants misheard or were unable to hear a specific word, the context provided by the rest of the sentence will have reduced the impact on their understanding of the sentence's meaning.

All ratings, other than those referring specifically to disruption caused by or to the MPC



visual content, exhibit no significant differences between the SD-V condition and at least one of the auditory displays. Particularly within the ratings of disruption to or by the MPC audio, one would expect that competition for attentional resources would more greatly affect the auditory SPC. It is possible that competition for linguistic processing resources led to participants feeling disruption between the MPC speech and reading the SPC (Wickens, 2008). This is supported by the comment from a participant stating that they “[r]eally felt the interference of the voices in the show with what I was trying to read on the phone” [P27, *unmodified*]. For a number of the scales, the results for the unmodified condition relative to the modified conditions appear more preferable (i.e., lower disruption or workload) for the SD-V group compared to the SPC source conditions. While interaction terms in these cases were not found to be significant, it is recognised that a more statistically powerful study may prove otherwise.

The participants’ ratings for the SD-A and SD-V treatments only differed significantly for disruption related to the visual content of the MPC. In both instances, the ratings of disruption were significantly higher for the SD-V presentation. This result is in keeping with multiple resource theory (Wickens, 1980, 1984). Auditory SPC would be expected to be associated with minimal interference, as the two sources require different attentional resources. Visual SPC and the MPC visual content, however, both need to be attended to visually and therefore interference between the two can be expected. Comparisons of the results for the ratings of disruption to the MPC visuals by the SPC and to the SPC by the MPC visuals appear to show a larger effect for the former. This suggests that participants’ generally felt that the MPC was being disrupted by the SPC rather than vice-versa. The feeling of disruption to the MPC visuals is illustrated in the comments made by the participants expressing a sense of missing out. The most well evidenced instance of this is comments about missing out on the scene with the jaguar, which many participants obviously wished to attend to visually.

From the qualitative analysis, one participant from the SD-A group spoke of visually attending to the phone, despite the lack of information on it. Inspection of videos of the trials indicated that participants glancing towards the secondary device was not uncommon in the SD-A condition. Some participants were also observed to glance towards the secondary device during the front and centre presentations. In the side condition, several participants appeared to turn their head towards the source, either following the notification sound or during the SPC presentation. In both front and side conditions, participants appeared to

look towards the source without turning their head. Given these observations, it would have been interesting to compare how much participants looked away from the screen during the SPC presentations. A formal analysis of this behaviour was not possible, however, due to the limitations of the captured footage which made it difficult to resolve where participants were looking. Difficulties were compounded by the lack of precise timing metadata with the videos, needed to identify exactly when visual SPC presentations began and ended.

Intriguingly, the results indicate that participants felt that the location of the ASPC source significantly impacted on the disruption that was caused to the visual elements of the MPC. The SD-A location was found to be significantly less disruptive than both the front and side conditions. The cause of this effect is far from clear. The higher presentation sound level may have meant that participants were more inclined to attend visually to the fixed sources than the handheld one. This is, however, not supported by the other data, as the comment about visually attending an auditory source comes from a participant in the SD-A condition and it is not possible to conduct a formal analysis of participants' visual attention from the videos of the experimental sessions.

The lack of a significant effect from the SPC source on participants' perceptions of physical demand is an interesting result. One might expect that the need to scroll through text in the visual condition, or participants' movements of the secondary devices in either of the two handheld conditions, would contribute to elevated ratings of physical demand. As participants in all SPC source conditions were required to hold the device or place it somewhere for each trial, all participants will have associated some physical demand with the displays. The physical demand ratings for all conditions, however, are extremely low. Close inspection of the aggregated ratings for each SPC source group does show the SD-V condition as having the highest trimmed-mean for physical demand. The difference is extremely small, however, and impossible to separate from random, between-group variation within this study. Any additional physical exertion from either of the handheld conditions would appear to be inconsequential in practice.

### **SPC information design**

The qualitative analysis demonstrates that preferences to the types of information presented as SPC varied between participants. It is clear that in a real world scenario, it is important to allow participants the choice to opt-out of content they are not interested in. The use of symbolic earcons for the SPC notifications, as suggested in Section 6.4.5, alongside an

interactive version of experience, may go some way to addressing this problem.

The stimuli used within this study for the SPC were fairly long at approximately 30 seconds in duration. In a real-world implementation, it is recommended that ASPC elements are of shorter durations, as this was highlighted as an issue by a couple of the participants. The provision of interactive control over the SPC may serve to reduce this requirement, as participants could simply stop the SPC whenever they wish to.

Despite the prevalence of comments referring to the issue of on-screen talkers amongst modified treatments, participants still considered the no-speech and atmosphere treatments to be most preferable. It is possible that participants felt that the added interactivity would enable them to avoid this audio-visual conflict. If a user were able to stop the SPC and quickly return to the original MPC mix, the amount of information missed in these scenarios could be reduced. Should dialogue removal be employed, the alternative suggestion of subtitles for the MPC, made by participants within the comments, could be an effective approach for reducing feelings of missing out and disruption. Providing subtitles could allow the user to skim-read or pick out basic information about what the MPC dialogue was talking about, yet still mostly focussing on the information of the SPC. While it is not expected that a user would be able to fully follow both streams as well as if they had been presented in isolation, the information from the MPC subtitles may be enough to alert them to something more interesting in the MPC so that they can switch back to it. A major benefit of this approach would be the ease with which it could be realised. As programmes are delivered with synchronised subtitle tracks anyway, it would be a simple case of allowing the device coordinating the experience to toggle them on or off based on the times when SPC was activated.

It is also possible that attenuation of the dialogue, rather than completely muting it, would have reduced the feelings of missing out. This approach would reduce the impact of a lack of interactivity, as participants would still be able to listen to MPC at a lower level. As discussed previously in Section 7.4.7, a difference in levels of the competing speech sources may prove beneficial to the understanding of both streams. Nevertheless, the presence of the MPC may still contribute to informational masking of the ASPC, or distract from VSPC. Also, as the MPC dialogue would be at a reduced level, it would suffer more from masking from ASPC. This could lead to more frustration, as the participant would be aware of the dialogue's presence but may struggle to understand it.

## Pausing

For the display methods tested here, the qualitative analysis showed that many participants were concerned about missing out on content, particularly from the main programme. Pausing the programme would avoid this problem when attending to SPC, albeit with some alternative compromises to the experience (as discussed in Section 6.4.2).

Though a pausing behaviour was not included as part of this experiment, there are some points arising from it that are likely to remain applicable. The complete muting of MPC was negatively viewed by the participants. Some of this was undoubtedly due to the main programme's continuation and concern about missing information. Participants' comments indicated, however, that muting also made the MPC and SPC streams feel separate and the inclusion of atmospheric sounds and music helped to avoid this. This would suggest that when pausing a MPC to display SPC, continuing some programme sounds would provide an improved, more integrated user-experience. Given an object-based broadcasting system, this could be implemented by either providing additional music and atmosphere tracks, or developing an intelligent looping system to select elements from the paused scene to continue the atmosphere. This kind of behaviour is commonplace within video games, when a user pauses gameplay to access a menu. As television experiences become more interactive, there is much to be learnt on technical architectures and design from the video game industry.

### 7.4.8 Limitations and future work

Due to the experimental set up, there were a number of compromises that had to be made in terms of ecological validity. Though attempts were made to make the experimental space feel like a normal living-room, this was only partially successful. Also, in the interests of experiment duration, clips were short, which may have meant participants did not engage with the programme in the same way that one might with a normal television experience. It is clear that a final prototype would require testing in more ecologically-valid settings to explore participant responses. On a related note, this experiment has investigated the experience of single users in isolation, as opposed to multi-user experiences. Interesting opportunities remain for exploring the impact of these experiences on multi-user scenarios.

As the television programmes used within the experiment had been broadcast previously, participants may have already experienced the MPC. Following the experiment, one participant indicated to the author that they had previously seen the programme used in

the experiment. This is not ideal, and an implicit issue associated with using released content. Using unreleased content is clearly preferable, though it was not possible in this case. For researchers, however, compromises must be sought between professional production, public exposure to the content, and the availability of media. Similarly, participants may have already had knowledge in the areas of the SPC prior to the experiment. In fact, one participant spoke of the SPC interrupting the MPC “to be told something I already knew” [P15]. While screening participants for prior experience could reduce the influence of this in future experiments, it may also have priming effects and encourage participants to seek information prior to the experiment. This is presented as a methodological consideration for future experiments of this nature.

As MPC treatment was a within-participants factor, it will have had more power to detect effects than the between-participants conditions or the interaction. Visual inspection of the data suggests that some interaction effects may have emerged if either a larger sample had been used or it had been possible to include SPC sources as a within-participant factor. Given the current number of factors, this would not have been feasible here due to the requirement for either an extremely large number of participants or an excessively long experiment duration for each participant ( $\approx$  2hrs15 without breaks). Considering the results of this experiment, however, it is felt that any future work can now focus on a smaller sub-set of conditions and adopt a more powerful experimental design.

The experimental analysis presented here has focused on clips taken from one programme. While attempts were made to phrase questions in a manner that did not directly relate to the content, it is likely that the content of the clips still had an effect on the participants’ responses. Further exploratory work would be required to explore how other programmes and genres would be affected by the MPC treatments considered here.

The behavioural analysis in this experiment was limited due to the method by which the session was recorded. The use of multiple camera angles or eye-tracking may have allowed classification of when a participant’s gaze shifted to the fixed source locations. K. Krejtz *et al.* (2012) and I. Krejtz *et al.* (2012) found that the addition of AD altered the way in which children visually attended to the video. It would be interesting to explore whether effects of this type are also present with the types of presentation considered here.

Some aspects of this study appear to have implications for experiences that pause the main programme. Further testing would be needed to confirm these and explore the additional complications that this raises.

As the focus of this experiment was on audio-visual MPC with participants reporting normal vision, it is difficult to determine how such treatments would be considered by users who have visual impairments or are blind. With less severe visual impairments, where a participant is still able to determine most of the on-screen activity, it seems likely that some of the same factors will apply. For example, such users will still be able to see when a person is on-screen that they should be able to hear talking. It is likely that influences from the on-screen action will have less effect as visual impairments become more severe. It is possible in these cases, therefore, that extreme lateral separations (e.g., the side location in this study) would become more preferable, as the user will be less inclined to orient their attention towards the screen. In such scenarios, new challenges are also introduced. Firstly, removing programme sound could have a greater impact on the user's understanding of the programme's narrative, because of the lack of visual information from the screen during the SPC presentations. Furthermore, it is likely that a user with severe visual impairment or who is blind will also wish to watch the programme with audio description turned on. This additional speech source may lead to confusion regarding whether voices belong to AD, MPC dialogue or SPC. Further research must be performed to determine the specific requirements for such cases. This is left as an area for future researchers to consider.

## 7.5 Summary

This chapter has presented an experiment comparing different methods for presenting auditory and visual SPC with MPC modifications. This required the development of a prototype system capable of simulating these experiences. Auditory presentations from fixed locations ( $0^\circ$  and  $90^\circ$ ) and a handheld location (smartphone) were tested and compared to a visual presentation of text on a smartphone screen. The effects of presenting the SPC alongside the MPC were investigated when the MPC soundtrack: was left unmodified; had the centre channel (containing all of the dialogue) removed (no-speech); was replaced with a 5.1 upmix of the atmosphere and Foley (atmosphere); or was completely removed (mute).

The work in this area has required the development of a novel methodology, which has been outlined. Analysis included the statistical analyses of participant ratings for disruption, workload and preference, plus a thematic analysis of comments. Results have indicated that a presentation from a secondary device with the dialogue removed from the MPC is the best option out of the auditory options tested here. This approach appears to compare favourably to the visual presentation of SPC with the only significant differences being in

relation to disruption to and from the MPC visual content, which were lower for the auditory presentation.

Additional elements to test have been identified. Now that a large study has been completed, it is recommended that more focussed studies are conducted to investigate aspects of the data that may have been obscured by low statistical power. A number of potential extensions have also been suggested as a result of participant comments, including the addition of subtitles to the experience.





# Chapter 8

## Conclusion

### 8.1 Introduction

This chapter considers the work that has been presented throughout this thesis. Both experiments are discussed with regards to the hypothesis that was presented in Chapter 1 and further consideration is given to the potential for concurrency within consumer auditory displays, drawing on the practical and theoretical elements of this work. Directions for future work are identified that may offer further insight on elements that have been uncovered in the course of this project.

Chapter 1 introduces the problem of visio-centric design in connected television interfaces. It emphasises the importance of the development of non-visual experiences both for users who have special access needs, and for users who are unable or unwilling to attend to a screen for other reasons. Audio is identified as an alternative modality to consider for these experiences and the potential of concurrent auditory display is highlighted.

This project has come at a time where traditional broadcast models of television are being challenged by internet delivery and new devices are allowing television content to be experienced in new contexts. To facilitate better design of interfaces and television services, there is a great deal of work to be done in understanding the ways in which these new technologies are being exploited by users. Chapter 2 introduces the current state-of-the-art in television user experiences and identifies areas that appear likely to be significant in the near future. Terminologies are introduced for the discussion of these experiences, including the taxonomy presented in (Hoare & Hinde, 2016), which the author collaborated on for this project. It is hoped that this will serve to aid other researchers in the discussion of these

experiences and also provide a set of factors to consider for practitioners looking to create new television experiences.

In the discussion of future television experiences in Chapter 2, a particular emphasis is placed on the importance of object-based delivery as a means to facilitate adaptable experiences for television. This discussion highlights menu navigation and companion experiences as use cases that are likely to be important in future television experiences.

In order to design effective auditory displays, it is important to understand factors which govern human perception of sound. Chapter 3 introduces the physiology of the human auditory system and key psychoacoustic principles which affect how we perceive sound. Our hearing systems have to deal almost constantly with multiple acoustic sources. The features that allow us to discern between different sources in these mixtures are given particularly close attention due to their significance in the design of successful auditory displays that comprise concurrent sound elements. Speech is introduced and the factors that are important within multi-talker scenarios are reviewed in detail.

Following this, Chapter 4 reviews auditory display methodologies, including both non-speech and speech based displays. Issues with the use of serial speech interfaces are highlighted and the limitations of non-speech methods are noted. A particular focus is given to displays that have been developed with concurrent audio streams of either speech or non-speech. Auditory display concepts are discussed within the context of television use cases. Serial speech is identified as the default solution for providing auditory displays, but problems are identified in terms of the amount of time it requires to present information. Non-speech methods provide a range of techniques for representing information through other auditory codes. All of the current methods either require users to learn links between items and their auditory representation (e.g., earcons and auditory icons) or are ill-suited for representing the type of information found within television interfaces alone (e.g., sonification and audification). Concurrent speech is highlighted as a method that appears to have the potential to provide the benefits of serial speech interfaces, but also to allow faster and more timely presentation of information. The potential for applying concurrency in television menu navigation and orchestrated synchronous companion experiences is discussed.

## 8.2 Menus

Chapter 5 explores the use of onset asynchrony in spoken menus. Results reported by Ikei *et al.* (2006) show very high accuracy rates from a spatial auditory display using onset asynchrony. While there is an increase in presentation speed from these displays, from the work presented by Ikei *et al.* (2006) it is unclear what affect concurrent displays with onset asynchrony have on the speed of navigation tasks. Furthermore, it is unclear whether it is the onset asynchrony or the amount of overlap that is the critical factor.

To investigate these factors, a menu display was designed (Section 5.2). The virtual display presented three spoken sources at a time, which were distinguished by their occupying different lateral locations, having different fundamental frequencies, and being presented in a particular order with onset asynchrony. Sliding window and segmented interaction models are presented and discussed. The removal of display redundancy and the reduction in the amount that sources move justify the use of the segmented model. Additionally, the interface allowed users to select any of the three currently presented items at any point. By incorporating these design features, the display was optimised for fast and efficient navigation.

In order to evaluate the impact of variations of onset asynchrony it was necessary to develop a suitable methodology. The pilot study presented in Section 5.3 was a first attempt at this. It uncovered several important methodological considerations, which led to an updated methodology that was used for the main experiment. The lessons learnt from the issues in the pilot, and the final methodology used in the main experiment should inform future researchers seeking to evaluate performance in menu navigation tasks.

The experiment, presented in Chapter 5, recorded participants' ability to navigate to and select a given target word with different onset asynchronies and lengths of words. Results indicate that onset asynchrony appears to be a good descriptor for the effects observed on task duration and workload. The error-rates, however, appear to have been affected by both overlap and onset asynchrony. The results suggest that an optimum onset asynchrony exists around 380 ms. For the short words, this condition led to significantly lower error rates than when the longer words were used. The fact that with this degree of asynchrony the short words did not overlap suggests that the use of shorter words or temporally compressed words, presented with a short inter-stimuli interval, is a better approach than allowing the stimuli to overlap. This finding should inform designers considering the design of auditory menu displays based on overlapping speech.

It is possible that the choice of length and of the phonetic nature of the stimuli in the experiment affected the ease with which the display could be used. It is thought likely that use of a real-world collection of options would increase its ease of use. Giving the users a definite target to identify made their task easier than would have been the case if they had been required to browse the menu, as occurs with programme selection menus (Elsweiler *et al.*, 2010). From this, it seems unlikely that concurrency in menu systems offers an advantage over fast serial speech displays.

The functionality to repeat triplets and to navigate back in the list was removed from the prototype system used in the experiment. The removal will have caused more errors to be observed than would have been the case if the full interface had been used. It is possible that including these features may have facilitated shorter, more accurate navigations in some concurrent conditions. As long as the time taken to repeat the triplets or return to the previous triplet reduces the presentation time compared with the time needed to present all of the items in isolation, task duration and error rates would be reduced. These restrictions were, however, necessary from an experimental perspective. Providing the ability to navigate back through the menu can cause a large positive skew to task duration data, as participants can navigate back and forth in the list repeatedly missing the target. In the pilot study, the repeating function was found to lead to disparities between the ways in which participants used the interfaces. Some made frequent use of the function, while others used it only occasionally. These behaviours would have created bimodal distributions of navigation time, which would have made statistical analysis problematic.

While the performance metrics pointed towards a non-overlapping condition as being favourable, the feedback from the participants in the post-experiment interview was less clear cut, with preferences for overlapping and non-overlapping displays both being expressed. It is possible that this is in part due to participants (a) being unaware of which specific conditions were overlapping and (b) not being informed of any errors that they made. The way in which the experiment was structured makes it very difficult to draw robust conclusions about the users' perceptions of the experience with the different onset asynchronies and overlaps. The analysis of concurrent menus presented here has concentrated on performance metrics (accuracy and navigation speed) with some subjective workload ratings. The limitations of the methodology used to collect feedback means that very little can be concluded about the experiential aspects of the individual displays. It may be, therefore, that the displays which are preferred do not actually provide the best performance or lowest workload metrics.

Within the context of a consumer display, it is important that a display should provide a good compromise between providing satisfactory speed and accuracy performance, and creating an interface that is pleasant to use.

### 8.3 Orchestrated synchronous companion experiences

Chapter 6 presents a review of literature on the effects of attending to speech while reading, which suggests that interference can occur between the two tasks, and that users are unlikely to be attending to both streams concurrently. This finding and the success of VSPC suggest that the disruption that would be expected with the addition of ASPC may not be excessively problematic. The design of a system is described for the presentation of orchestrated synchronous companion experiences using audio. Suggestions are made for an interaction model and presentation methods that can facilitate individual components within shared experiences.

The design proposed in Chapter 6 is then used as the basis of an experiment in Chapter 7. A difficult issue was faced in determining the best method by which to evaluate experiences of this type. As the experiences are novel, standard methodologies have not yet emerged for comparing different variations. A novel methodology is presented in which participants are asked to rate the amount of disruption from the SPC to the experiences of the individual audio and video components of the MPC, and to the MPC as a whole. Participants were also asked to rate the amount of disruption by the MPC and its audio and video components to the experience of the SPC. To evaluate the workload, participants rated the mental demand, temporal demand, physical demand, effort and annoyance that they experienced during each of the experimental conditions. An additional rating indicated participants' preferences to each of the displays. In addition to these quantitative methods, participants were also asked for comments on each of the displays which they experienced. A pilot study, presented in Section 7.3.5, helped to tune this methodology prior to the main experiment.

A criticism of this approach may be that the assessment relies too heavily on subjective ratings of experience factors and qualitative feedback, as opposed to objective performance measures. After considering the alternatives and the feedback from the qualitative data, no objective measurements have been identified that seem likely to provide meaningful insight into these experiences. Eye-tracking has been identified as an element that would be interesting to measure, to see what effect different SPC displays have on visual attention. It is, however,

unclear what the desired response would be. While a participant who directs their gaze away from the screen may be considered to be a negative indicator, it is less clear what to conclude if different users gaze at different regions of the display.

The experiment compared participants' responses to different presentations of SPC in the form of audio from different locations and as text from a portable device. These were investigated for a variety of MPC treatments, involving the removal of different elements of the MPC soundtrack. The results of the experiment indicate that, of the MPC treatments, the one in which only the centre channel containing all of the spoken material was removed led to less disruption than the other muted and unmodified treatments. Furthermore, this treatment did not lead to higher ratings of workload than muting and it was preferred. Fewer significant effects were observed between the SPC source conditions, which is believed to have been partially due to the design of the experiment.

Of the auditory display conditions, the treatment in which the secondary content was presented from the phone (SD-A) appears to have been the best. It was rated significantly lower than the side condition in terms of the amount of disruption it caused to the experience of the MPC visual content and the experience of the MPC as a whole, and of the mental demand imposed. It was also rated as less disruptive to the visual MPC content than the frontal source. The SD-A condition compares favourably with the visual equivalent (SD-V) with the only significant differences being lower disruption ratings for the SD-A presentation caused to or by the MPC visuals. These results suggest that auditory display of SPC from a secondary device is a feasible alternative to VSPC presentations and can even prove advantageous.

The low ratings for the SPC side condition are somewhat surprising, as one might expect the spatial release from masking to have benefitted the users. This interesting finding challenges the efficacy of laterally positioning auditory sources when users are directing their visual attention forwards. Support for this view comes from Spence & Read (2003), who conducted an experiment which indicated that participants were less negatively affected by having to perform a driving task while repeating an attended stream of speech when the target speech was presented from the front than when it was presented from a lateral location. This may suggest that spatial separations cannot be used in auditory displays where a user is actively visually attending to a source at a specific location. Where concurrent auditory displays are concerned, however, there may be an arrangement with smaller angular separations that can compromise between SRM and the difficulty caused by the audio-visual links in spatial

attention.

From the thematic analysis of participant comments, it is clear that the presence of on-screen talkers when their voices have been removed is a large issue with the treatments considered in this study. Alternative presentation methods exploiting object-based broadcasting principles are suggested, but it is clear that further work would be required to find a satisfactory solution. An alternative solution, which was formulated from participants' comments, is the use of subtitles for the muted talkers. Additional testing would be required to explore the impact of this.

To simplify the experimental design, participants were not given the facility to choose whether or not a SPC element was presented (as proposed in the design section 6.4.4). The lack of this feature detrimentally affected the usability of the displays used in the experiment. The option to decide if SPC elements are played is considered to be an important feature that would be beneficial in presentations of ASPC and VSPC where the MPC audio has been modified.

This work on the presentation of ASPC as part of scheduled orchestrated companion experiences represents a new area of research. The experimental work on this topic presented in this thesis provides a basis for later research, both in terms of the methodology that has been used and the findings. It is also hoped that discussion of the design of these experiences will also serve to inform those in industry who are considering the design of synchronous companion experiences for television or other media.

## 8.4 Concurrency within consumer interfaces

The results from the two experiments have implications that extend beyond their respective use cases. Both show that the presentation of concurrent speech streams is problematic within user interfaces. They suggest that optimum conditions should involve no overlapping speech. While presenting multiple speech sources concurrently increases the speed with which information can be presented, the interference between the streams and their reliance on the same attentional resources appears to place severe limitations on how useful these displays are. The results of the second study, however, point to the experiential benefits that can arise from providing some concurrency within user experiences. Concurrency can serve to smooth the transitions between different interface states or informational sources. This is a useful consideration for those designing auditory displays who wish to provide an experience which

feels integrated, even when it comprises disparate parts.

The two designs proposed in this thesis are very different. This is a reflection of the wide variety of ways in which concurrency may be used in auditory displays. The menu display represented a case in which the user was given a specific piece of information which they had to locate within the display. Though cues were present in the display to aid stream segregation, participants did not know which stream would contain the information of interest. Conversely, within the companion experience, participants were not given any specific instruction about which stream they should attend to. The streams had constant characteristics (i.e., voice and spatial location) and so participants will have had a strong awareness of the cues associated with the stream that they wished to attend to. The informational density of the two experiments also varied considerably. The stimuli used in the first experiment were very short words and contained all of the critical information within one part of the word. By contrast, the second experiment used long extracts as stimuli.

The findings of this work may apply to other consumer use cases in which similar display issues are confronted. While the work on menu design focussed on television use, it may be applied to menus in a host of other scenarios in which users are not able to visually attend to a screen (e.g., in-car entertainment systems and portable media players). Taken at the most abstract level, the results of this study may have implications for any scenario in which users must choose from one of several short spoken options.

The orchestrated companion experience findings may also have applications in radio, although further consideration would be needed to account for the different auditory elements and the lack of clear locations that visual attention is likely to be directed towards. The types of information represented by SPC and AD are different and have been considered separately within this thesis. It is possible, however, that some of the display methods considered here, combined with object-based delivery, could provide improved experiences for some AD users by facilitating variable length descriptions rather than the pausing methods that other authors have proposed (Chapdelaine & Gagnon, 2009; Encelle *et al.*, 2013).

The focus of this work has been on the connected television use case, in which participants consume television content within a traditional living room context. Different interaction models will be required for different devices and use contexts, and identifying the best ways for users to interact with audio is an open research question. The second study found that providing users with tangible sources is advantageous. It is not clear to what degree this was due to proprioception facilitating spatial attention or to users having dynamic control over the



level of the ASPC. Either way, designing displays in which users can physically manipulate sound sources seems like an interesting method which can be facilitated by the IOT and which has the potential to improve user experiences with concurrent auditory displays.

## 8.5 Implications for the hypothesis

The hypothesis presented in Section 1.4 states that: concurrent auditory displays can:

1. facilitate faster navigation of menus without negatively affecting accuracy or user experience;
2. provide less disruptive and demanding display of additional secondary program content than serial alternatives.

Chapter 5 explores the first part of the hypothesis. The experiment found that the optimum asynchrony for navigation times appears to be around 280 ms, which led to some overlap with both groups of words, short and long. However, the shorter words with a short gap between them provide significant benefits in terms of error rates when compared to either the use of a shorter onset asynchrony so that words were overlapping, or the use of the same onset asynchrony with longer words which therefore overlapped. While there was a small increase in average navigation time from the optimal 280 ms, it was not statistically significant. Considering the impact of higher error rates on navigation times in real-world interfaces, it would seem that this form of display would actually be faster to use. Furthermore, by providing participants with a defined target, this task was considerably easier than that which users would be expected to perform in programme choice menus. The effects of concurrency are expected to be more severe in these scenarios, as users will have to understand all of the items in order to successfully pick the best one. This result means that we are unable to reject the null hypothesis that concurrent auditory displays do not facilitate faster navigation times without significantly affecting accuracy or user experience.

The results from the second study provide partial support for the second part of the hypothesis. While concurrent presentation of speech was not found to be significantly more disruptive than the muted conditions, it was considered to result in considerably higher mental load and effort, and led to the experience feeling more rushed. The no-speech condition was found to be less disruptive than the muted condition in terms of the disruption to the MPC soundtrack and MPC as a whole by the SPC. This condition was not found to be significantly

different to the muted condition in any of the workload ratings and was rated significantly higher in terms of preference compared to all other conditions. It is therefore concluded that concurrent auditory display can provide a less disruptive display of additional secondary programme content that is no more disruptive than serial alternatives when only one of the concurrent streams comprises speech.

## 8.6 Further work

Chapter 7 outlines some areas for further work regarding the orchestrated synchronous companion experiences, which for the sake of brevity will not be re-iterated here.

The problem of how to represent large, dynamic menus most efficiently remains unanswered. While this work suggests a grouped display using short or temporally compressed words, this has not been verified with a larger variety of words or compared to alternatives, such as one-at-a-time serial speech interfaces. Though some work has been done previously in this area by Sodnik *et al.* (2011), much larger inter-stimuli intervals were used, which may have affected the results. Further experimental work is therefore needed here to settle on the optimal design.

This thesis has focussed on the use of concurrency within two specific use cases. It is possible that other areas may benefit from concurrent displays. The work by Guerreiro & Gonçalves (2014, 2016) suggests that concurrent speech can prove useful when scanning for information from several spoken sources. Additional work is clearly needed to identify why benefits are not observed in some cases, such as those presented in this thesis. At the moment, conflicting results make it difficult for interface creators to place confidence in producing displays that exploit concurrency.

Guerreiro & Gonçalves (2016) noted that greater performance benefits may occur for users who have more experience in using concurrent displays. Again, though some training has been provided in both studies in this thesis, it is clearly not equivalent to the experience gained by someone who has used an interface for several days. As far as the amount of concurrency in menus is concerned, it may be that this is something that can slowly be increased as participants become more used to an interface. The effects of long-term usage of these types of interfaces also require investigation. Furthermore, the studies have been very tightly restricted in terms of the interactions that are allowed and what tasks the displays are used for. It is recommended that longitudinal studies of concurrent interfaces may help

to provide a more robust insight into the experiences that these displays facilitate.

These experiments have focused on samples of general population. There are clearly groups of users who would find these systems difficult to access, such as those with hearing damage or conditions affecting selective attention. While pre-existing visual interfaces may satisfy the access needs of these users in some cases, it may be that alternative auditory representations are more appropriate. Further consideration of these is needed.

Both experiments presented within this work explored the effects of different degrees of concurrency with very restricted sets of stimuli. Further experimentation with other stimuli sets would be required to add weight to the findings of these studies and to confirm that observed effects are not isolated to the test materials used here. Also, in both cases, trends were noted that failed to reach significance. Repeat studies with more participants, or more focused experimental designs are recommended to shed further light on these observations.

The second study, which considered scheduled, fixed, orchestrated companion experiences, begins to provide methodological insights in addition to those concerning the design of displays. Quantifying user experiences in orchestrated companion experiences is methodologically challenging and further work is needed by researchers and practitioners to define measures by which these experiences may be compared. Studying auditory attention is currently a challenge that is beyond the reach of most researchers and practitioners considering the design of auditory displays. Being able to measure what users are attending to in these complex auditory scenarios could facilitate the design of better auditory displays that remove unnecessary or distracting elements. From these findings, additional possibilities may also open up, such as the creation of attention-aware displays, which may be able to reduce the amount of physical interaction that these systems require.

The findings of the second study may have relevance to other categories of synchronous experience. It seems highly likely that all orchestrated companion experiences could benefit from the insights of the study presented here. It is interesting also to consider how the display of non-companion or unrelated content may be modified so as to facilitate improved synchronous user experiences. This could take the form of an application which alters the output of a secondary device when the user watches a television programme. (Examples include presenting an email or a related article that you have found as sound and modifying the programme soundtrack as you continue to watch the programme.) Further research would be necessary with these variations to confirm that the findings are, indeed transferable. The display of unrelated content seems likely to be most problematic, due to the probability that

the user's motivations differ to those associated with synchronous experiences comprising related content.

This thesis has only considered the application of auditory display principles in two use cases within connected television user experiences. It is clear from the discussion presented in Chapter 1 that many more experiences merit consideration. Additionally, over the course of this project, virtual reality (VR) technologies have emerged as an ever-more feasible prospect within audio-visual storytelling. These systems may benefit even more from auditory displays as a means of providing additional information without breaking immersion within the visual scene by overlaying additional information or menu structures. As these experiences become more common, the potential of audio within VR interaction should be considered.

Finally, many opportunities exist beyond the non-visual displays considered here. Haptic, olfactory, and taste displays should also be investigated further, as they may create television experiences which are more accessible, entertaining and immersive. It is important that, as new television experiences are developed, designers depart from the visio-centric tradition, to ensure better usability for all, and new and exciting opportunities for user experiences are not over-looked.

# Appendix A

## Publications

### A.1 Onset Asynchrony in Spoken Menus

Paper has been resized to fit the margins of this work. The original can be accessed at:  
<http://hdl.handle.net/1853/54112>

## ONSET ASYNCHRONY IN SPOKEN MENUS

*Alistair F. Hinde, Michael Evans*

BBC Research and Development,  
5 Dock House, MediaCity UK,  
Salford, M50 2LH  
United Kingdom  
afh508@york.ac.uk,  
michael.evans@bbc.co.uk

*Anthony I. Tew, David M. Howard*

Audio Lab, Dept. Electronics,  
University of York,  
York, YO10 5DD  
United Kingdom  
tony.tew@york.ac.uk,  
david.howard@york.ac.uk

## ABSTRACT

The menu is an important interface component, which appears unlikely to be completely superseded by modern search-based approaches. For someone who is unable to attend a screen visually, however, alternative non-visual menu formats are often problematic. A display is developed in which multiple concurrent words are presented with different amounts of onset asynchrony. The effect of different amounts of asynchrony and word length on task durations, accuracy and workload are explored. It is found that total task duration is significantly affected by both onset asynchrony and word duration. Error rates are significantly affected by both onset asynchrony, word length and their interaction, whilst subjective workload scores are only significantly affected by onset asynchrony. Overall, the results appear to suggest that the best compromise between accuracy, workload and speed may be achieved through presenting shorter or temporally-compressed words with a short inter-stimuli interval.

## 1. INTRODUCTION

The menu is a common feature deployed in user interfaces to allow users to navigate and find content of interest. With the development of ever more sophisticated search algorithms, it may seem as though the menu's role in user interfaces is soon to be confined to legacy software. Search functions, however, work best with well defined target items. In interfaces for entertainment systems, such as the electronic-programme-guide (EPG) on televisions, users may have only loose criteria governing their search (e.g. they may be wanting to view a comedy or plan their evening's viewing [1]). Within browsing scenarios such as this, search functions may, at best, reduce the size of the menu structure that must be traversed, as it is highly likely that users will still have to navigate lists of possible matches. It is also worth noting that these browsing activities additionally expose users to information about alternative, or new content.

---

This work was conducted by BBC Research & Development and the University of York, funded by an EPSRC Industrial CASE award.



This work is licensed under Creative Commons Attribution Non-Commercial 4.0 International License. The full terms of the License are available at <http://creativecommons.org/licenses/by-nc/4.0>

Menu structures can be large and confusing to interact with and they pose a particular problem for people who are unable to attend the screen visually. Due to these complexities, non-visual representation of menus is an area of auditory display which has attracted a great deal of research activity. Traditionally, menus are represented using text-to-speech (TTS) rendering, as is found in commercially available screen readers or in telecommunications. This approach is effective as it allows large amounts of complex and novel information to be displayed. Whilst speech is undoubtedly a logical representation for textual information, temporal redundancy means that there are many situations in which it is not necessarily the quickest method of communicating information. Researchers looking for speed improvements have typically turned to non-speech methods, which aim to represent the information through instrumental motifs [2], ecological sounds [3] or sped-up speech [4]. These approaches have proven to be effective to varying degrees (compared in [4]), but where information is regularly updated and often unfamiliar, as in the case of an EPG, or where a large amount of information needs to be represented, the usability of such systems is likely to be severely affected and any advantages in terms of navigational speed could be lost. In such systems, the redundancy in speech could be a distinct advantage for reducing confusion.

The problem of how to make speech interfaces faster to use is well known and a popular solution is simply to increase the rate of the speech in the system, as regular users of screen readers are commonly reported as doing (e.g. [5]). However, an alternative solution is to increase the amount of information available to the user by using several talkers at once. This form of display, referred to as a concurrent-speech, or multi-talker display, is more analogous to the manner in which visual displays are used, in that a large amount of information is displayed to the user, who then chooses to attend to the item which they are interested in. To facilitate this behaviour, however, the concurrent speech must be separable as individual perceptual streams by the auditory system [6].

As the amount of information presented to a user at any moment increases, one would expect a user to have to work harder to focus on the desired stream of speech. Much of the work on the use of multiple streams of concurrent speech has been in the field of military communications (e.g. [7, 8, 9]), in which high workloads may be necessary to ensure that time critical information is received. By contrast, consumer menu interfaces are likely to have a lower threshold for what constitutes an acceptable amount of effort. Nevertheless, several authors have proposed auditory displays using concurrent speech for menu navigation in consumer

HCI applications.

Frauenberger and Stockman proposed a design using concurrent speech to navigate auditory menus using the idea of a virtual horizontal dial with items located around its perimeter [10]. The display used a virtual room with the centre of the dial positioned outside so that a maximum of three items from the menu would be inside the room, and therefore audible, at any time, along with two additional ‘preview sources’ if the selected item was a sub-menu [11]. The user navigated the menu by rotating the virtual ring using a game pad dial until the desired item was directly in front. The display made use of different voice identities and talking styles (i.e. voiced or whispered) to reduce between-stream confusions. When compared with the performance of a traditional screen reader, navigation times were found initially to be faster but in the later trials participants became faster with the screen reader. This was attributed to fatigue effects caused by the repetitive presentation within the prototype display.

Ikei *et al.* [12] proposed the use of multiplexed speech, where delays were included between the onset of successive spoken menu items. Speech sources were spatialised and a range of onset intervals, between zero and 500 ms, were assessed. Trials consisted of between two and four sources with different source orderings (i.e. from left to right or alternating between sources from either hemisphere) and either with or without a linear increase in attenuation applied over the course of each word. The work examined the impact of these display variations on participants’ ability to identify the temporal or spatial location of a target word within the mixture. It was found that with three or more voices high accuracy ( $\geq 99.7\%$ ) could be achieved with onset delays of 200 ms, however, this increased to 300 ms if adjacent sources were used and no attenuation applied. As noted by Ikei *et al.* [12], the optimal asynchrony without attenuation (300 ms) is about half of the duration of the stimuli (530 - 600 ms). This is important when considering the playback of more than two voices because when the onset asynchrony is less than half the duration of the stimuli, all three items overlap, but when the onset asynchrony is half the duration or greater, a maximum of two items are presented at any one time (Fig. 1). This simplifies the task in terms of both attentional load and signal-to-noise ratio. If the optimal asynchrony was purely due to this effect, a similar behaviour would not be expected within the two-talker condition. The two-talker results appear to indicate improving performance with increasing onset asynchrony, but ceiling effects in the participants’ performances make it difficult to ascertain the strength of this effect.

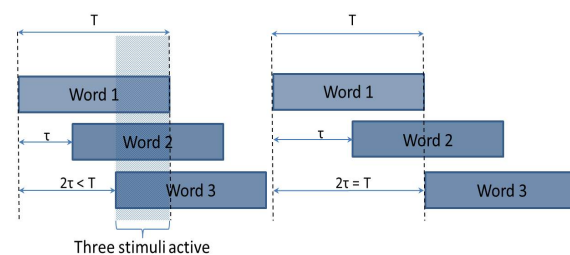


Figure 1: The reduction in the number of concurrent sources when an onset asynchrony of 50% or greater is used (adapted from [12]).

In his thesis, Parente proposed and tested an auditory display system using spatialised concurrent speech for computer-based GUI tasks [13]. The display consisted of five speech sources with differing vocal characteristics (accent, sex, identity), each responsible for reporting different types of information. In addition to this, speech was presented with a 200 ms onset asynchrony between streams. Testing, which compared task performance of the system to that with a conventional screen reader, showed that participants were able to navigate with reasonable accuracy. Interpretation of task durations was complicated due to the different interaction capabilities of the different displays (i.e. availability of search functions) and therefore it is difficult to determine what degree of advantage was provided through the use of concurrent speech.

In scenarios in which content is displayed to the user, who then elects to attend a particular stream (e.g. [14]), it would appear indisputable that concurrency has the potential to reduce presentation times. When users are expected to interact with concurrent speech displays, however, a reduction in task time is not a foregone conclusion. Users may take longer to respond due to the additional processing required to disentangle the contents of the display, or may be more likely to make mistakes due to reduced intelligibility. Therefore, there is still some uncertainty over the amount of time saving which these displays are able to provide. The usage of onset asynchrony alongside pitch and spatial separation of speech streams appears to be promising for reducing navigation times. Due to the inherent trade-off between improving response accuracy [12] and increasing overall presentation time, onset asynchrony is a factor which needs further consideration regarding its impact on overall navigation times. In addition to this, it is still unclear whether a display designer interested in the use of multi-talker display should specify onset asynchrony or overlap. The aim of this paper is to provide deeper insight into the effects of overlap and onset asynchrony in multi-talker menu displays with a particular focus on navigation time.

## 2. DISPLAY DESIGN

With a concurrent auditory presentation of speech, it is clear that users find it harder to identify and detect accurately the displayed words as the number of concurrent speech streams is increased [15, 9]. Ikei *et al.* [12] found that, when onset asynchrony is introduced into the display, very high accuracy can be maintained for greater numbers of talkers. The greater the number of concurrent talkers, the larger the potential saving in terms of navigation times. To investigate these issues a three-talker design was used in this study.

Many previous auditory displays using concurrent speech have made use of binaural processing to spatialise the audio sources (e.g. [10, 12]). For two reasons, it was decided to use intensity panning to lateralise the stimuli in this study. Firstly, when only two or three concurrent speech streams are used, it is not clear what advantage binaural spatialisation provides over intensity-panning-based lateralisation, as with intensity panning it is possible to confine sources entirely to one channel and therefore remove the energetic masking caused by sound arriving at the contralateral ear. Secondly, binaural spatialisation introduces perceptual factors which tend to vary between participants (e.g. externalisation and front-back confusions), whereas intensity-panning does not.

The display was designed to minimise the amount of interactions required. It is based on the idea of allowing the user to select between three items at a time or move on to the next set of three

items (referred to henceforth as a *triplet*) (see Fig. 2). This method was chosen as it reduces the number of interactions required compared to a display in which the target must be at a specific location for selection, effectively forcing the user to move through the list one item-at-a-time, as in [10].

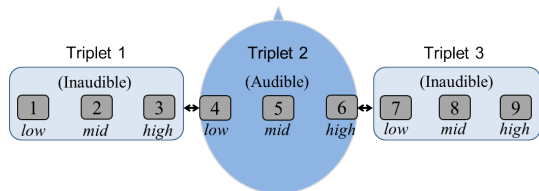


Figure 2: Illustration of display design concept (italicised writing refers to stimuli pitch).

The sources within a triplet were presented from maximally separated lateral positions using intensity panning, such that one source would appear on either side of the head through being presented to only the ipsilateral ear, whilst the third source was presented at the same level in both channels so as to appear in the centre of the head. Within a triplet, sources were also distinguished by a pitch difference such that the stimulus on the left was lowest and the stimulus on the right was highest. The order of presentation was kept constant and ran from left to right to correspond with normal reading direction. Whilst other studies have found advantages in modifying the talker’s apparent sex through vocal tract length modification [16], this was not attempted in this study to avoid making the voices sound excessively unnatural.

The system was controlled using five of the drum pads and one rotary dial on a USB MIDI controller. Two rows of pads on the device were assigned to different functions. Playback controls (start playback and navigation to the next triplet) were provided using two pads on the top row and selection of the three items in the currently selected triplet was performed with three pads on the bottom row. An earlier iteration of the prototype also allowed the user to navigate back in the list to a previously heard triplet, and to repeat the current triplet. This functionality was subsequently removed to minimise variance in the navigation time data by ensuring that users took the most direct root to the target. As a result of this, more errors were expected to occur than if these functions had been left in.

The prototype system was developed in Pure Data (Pd-extended 0.43.4) [17] and run on Ubuntu 12.04 LTS with a low latency kernel. To reduce computational load and the possibility of error the triplets were processed in advance, as discussed in Section 3.1. Therefore, the patch was mainly responsible for receiving the MIDI messages from the controller, handling trial configurations, file playback and recording responses.

### 3. EXPERIMENT

Sixteen participants were recruited from amongst the BBC Future Media department staff and its visitors. Volunteers who reported hearing impairments were not included in the study but no audiometric testing was performed on the participants. No attempt was made to recruit participants with vision disabilities, since the utility of a non-visual display is not solely restricted to people with visual impairments or who are blind.

During the experiment two participants experienced a fault in the software. For one of the participants this fault affected the experimental trials; therefore, this participant was excluded from the study and an additional participant was recruited to fill the space.

#### 3.1. Stimuli

The wordlists from the Modified Rhyme Test (MRT) [18] were used as the source of the words for the experiment. The MRT wordlists consist of two sets, each of which is made up of 25 lists of 6 words. Stimuli were selected from the first set, within which each list of words shares the same first consonant-vowel pair but differs in the final consonant (e.g. *page, pale*), or in some cases has no final consonant (e.g. *ray*). For this experiment 22 of the lists were chosen and for each of the lists one word was removed. Removals were mostly made because of the americanised pronunciations required for some words to share common vowel sounds (e.g. *pass, pat*), the lack of a final consonant, or because some words were deemed too unusual or inappropriate for the experiment.

The 110 words were recorded being spoken by a male talker, who was asked to enounce with minimal intonation and variation in word duration. Recordings were captured as 24-bit audio files with a sample rate of 44.1 kHz. Words were cropped to remove silences before and after they were spoken. Where possible, the crop was made at a zero crossing. In a few cases, however, no suitable zero crossing was available and the crop introduced a small discontinuity into the waveform. In such cases, the word was auditioned to ensure that no click could be heard. For some fricative consonants, low-level sounds at the start or end of the word were removed if they were considered not to be contributing to the intelligibility of the word. The durations of the stimuli were found to vary quite considerably, ranging from 301 to 709 ms with an average of 486 ms.

The stimuli were manipulated to have the same constant pitches and durations (either 360 or 600 ms) in Praat [19]. The pitch values of words in the centre of a triplet were adjusted to correspond to the average pitch of all of the stimuli. The pitches of the words on the right or the left in a triplet were adjusted approximately to plus or minus one ERB [20], respectively, compared to the pitch of the centre word. While other authors have reported benefits from much larger pitch differences [16], a comparatively small pitch difference was used here, as there was concern that further modification would have jeopardised intelligibility and may have led to users becoming distracted by the unnatural character of the voices. The durations were chosen to ensure that no word would be stretched to more than twice, or shortened to less than half, of its original length.

The stimuli were processed such that the words appeared to be approximately the same loudness in all onset asynchrony conditions. The onset asynchrony varies the amount of overlap between the words and this varies the overall loudness of the presentation. While it would have been possible to normalise the loudness of all presentations, this would have altered the levels of the words between presentation conditions. It was therefore decided that it would be more consistent to preserve the variations in overall presentation loudness. To achieve this effect the stimuli were mixed to equal loudness by ear.

The stimuli were combined into triplets and onset asynchronies were adjusted using MATLAB to ensure that they were as accurate as possible. Each triplet presented in the experiment met



the conditions that all words had to be from the same list, with the same word length and each word could only appear once within that triplet. This resulted in the creation of 10,560 triplets. The highest peak amplitude in the set of triplets was found and used to calculate the scaling factor necessary to bring this peak to an amplitude of (+/-) 0.9999 so as to maximise signal-to-noise ratio whilst avoiding any clipping distortion. This scaling factor was applied to all of the stimuli to ensure the relative loudness was not altered. The triplets were then exported as 44.1 kHz, 16-bit WAV files.

### 3.2. Procedure

The independent variables were onset asynchrony and word length. These variables had four [180, 280, 380, 480 ms] and two [360, 600 ms] levels respectively. The experiment was structured as a within-subjects design, with all participants experiencing all presentation conditions. Experimental trials were split into sessions of fixed word length. Each session contained four blocks in which the onset asynchrony was kept constant. This structure was imposed on the trials so that NASA TLX subjective workload assessments could be performed on each of the word length/onset asynchrony combinations. On completion of each block, a computer based version of the evaluation [21] was undertaken by each participant.

Each trial started with a target word being displayed on a screen and then, when the user was ready, they pressed the 'start' pad which immediately played the first triplet in the list. The user then would navigate until they found the target, whereupon they would select it by pressing the appropriate pad. In some conditions the target word was not present in the list, in which case the correct response was to navigate onwards from the final triplet.

Each list in the experiment was nine words long, with the target words, if present, only presented once at one location. In trials in which the target was present the lists were constrained such that each triplet had at least one stimulus that was not in the previous triplet; all words in a triplet were different; the target only appeared at the target location; all non-target words had to appear twice and never in the same lateral location. These constraints were put in place to ensure that there was variation between the triplets in a list and therefore avoid participants being able to rely on simply detecting that the triplet had changed. This led to 960 possible list combinations for each word at each location within the list. When the target was not included in the list only one item could appear three times, whilst all others would appear twice and words would never appear in the same lateral positions. These criteria led to 576 possible lists for each word.

The experiment was split into three sessions with twenty-minute breaks between them to reduce the effects of any fatigue. In the first (training) session participants completed an informed consent form and then were introduced to the system, an initial playback level was set (participants had the ability to adjust this throughout the experiment) and they were given practice tasks. During the training, each participant performed 40 tasks which consisted of 4 blocks of 10 trials, one block for each onset asynchrony. Each block consisted of all target locations and these were pseudorandomly allocated a word length condition such that 50% of each block was of each condition. Target locations were pseudorandomly ordered so that for each participant each target location could appear once for each trial number within a block. For each of the blocks the participant completed a NASA TLX questionnaire.

Participants were given additional guidance when it appeared they had not fully understood how to use the setup or were unclear on how to respond to the TLX questions.

The second two sessions consisted of the experimental trials (limited to a maximum of 20 minutes each), with each session containing one of the word-length conditions. Half of the participants were presented with the short words first and the other half heard the long words first. To reduce the influence of ordering on the onset asynchrony conditions over the training and experimental sessions, three counterbalanced Latin squares were used to vary the orderings. For each instance of the Latin square the dummy values were substituted pseudorandomly for onset asynchrony conditions, such that no dummy value represented the same onset asynchrony condition twice. A row from each of the Latin squares was then used for each of the sessions. The order in which the Latin squares were used for the sessions was varied for every four participants to produce variations for all 16 participants.

Within each block, target location order was varied pseudorandomly, with the restriction that for each participant each target location could appear no more than twice at each trial index in the training and experimental sessions and not at the same trial index as in the previous session. Each trial's list was randomly chosen from the 22 possibilities such that the same list was never used in two consecutive trials. The target word was randomly chosen from the list. The experiment list was then randomly chosen from all possible lists for which the target was at the specified location.

To ensure that data points were captured for all target locations, trial accuracy information was output from Pure Data and read by a Python script. This script analysed the configuration of the original trials and generated new repeat trials when a participant failed to select the target. These repeat trials were then added to the list of trials being read by Pure Data. Repeat trials were reordered and modified so that they used a different one of the 22 wordlists to both the preceding trial and the trial to be repeated. This ensured that the target identity and list were also different so that a participant's previous exposure to the same target location would have minimal effect on their performance. Two additional dummy trials were added using target locations for which the user had already registered an accurate response. These served as a buffer zone in the event that the participant's response to the final trial in a block was incorrect. As the Python script was editing the input to the Pure Data program while the participant was using it, the target location of the last output trial was used to decide where the repeated trials should be added. The repeats were added to the end of the original 10 trials until the last completed trial was beyond the eighth trial, at which point repeats were added after two trials. This ensured that no trials were altered after the user had already begun them and that a repeated trial was always separated from the original by at least two trials. For the repeats the list was randomly selected from the 960 (or 576 if for a 'no-target' trial) possibilities, making the chances of a trial sharing the same list as another trial negligible.

Following the experimental tasks participants were asked a series of questions regarding their experience with the interface (details of which are beyond the scope of this paper) and were then debriefed.

## 4. RESULTS

During the running of the experiment, on four occasions it was clear that the participant made several attempts to navigate to the

next triplet but had not pressed the pad with sufficient force, causing a significant delay in their navigation time. The experimenter flagged these trials during the experiment and repeats were generated, as if they had been incorrectly answered. Of the four affected trials, one had been a dummy trial. The original affected trials were removed from all subsequent analyses, with the data from the repeated trials being used in their place.

All statistical analysis was performed using SPSS. Further information on the statistical tests used can be found in [22].

#### 4.1. Total task duration

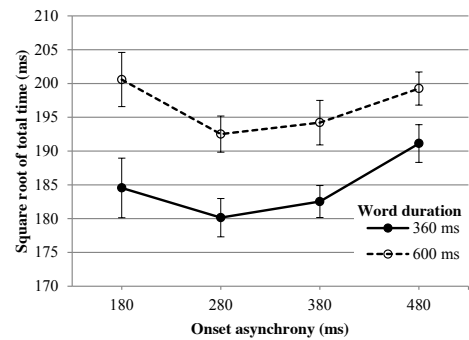
The duration of a trial (i.e. the ‘task’) was taken as the time from the playback of the first triplet, following the user pressing the ‘start’ button, to the time when a selection was registered by Pure Data. The task durations of all scoring trials were then summed over all target locations (including when the target was not present) within each onset asynchrony/word duration block for each participant. Trials in which the participant responded incorrectly or which were added as dummy trials were excluded from this sum. This effectively removed the nuisance variable ‘target location’ from the analysis, leaving each participant one total task duration for each experimental block.

A positive skew at one onset asynchrony/word duration combination was observed. Since this violated the normality assumption required for parametric analysis, the square root of the aggregated task duration data was used. A Shapiro-Wilk test confirmed that the transformed data was not significantly different from normal ( $p > .05$ ). A two-way repeated measures ANOVA (rm-ANOVA) was performed with onset asynchrony and word duration as independent variables.

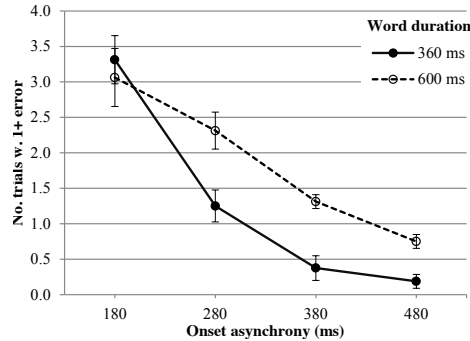
Mauchly’s test indicated that the sphericity assumption was violated for the onset asynchrony condition ( $\chi^2(5) = 13.5, p < .05$ ) and therefore it was decided to use the Greenhouse-Geisser correction ( $\epsilon = .60$ ). Sphericity was met for the word duration (2 levels) and the onset asynchrony  $\times$  word duration interaction ( $p > .05$ ). The results of the rm-ANOVA indicated significant effects for onset asynchrony ( $F(1.81, 27.1) = 8.79, p = .002, h_p^2 = .369$ ) and word duration ( $F(1, 15) = 25.3, p < .001, h_p^2 = .627$ ), while the interaction was found to be non-significant ( $F(3, 45) = 1.19, p = .323, h_p^2 = .074$ ) (see Figure 3a). *Post-hoc* pairwise tests were performed for onset asynchrony using a Bonferroni correction, which indicated that the 280 ms onset asynchrony conditions led to significantly shorter total task durations than the 180 ms condition ( $p = .038$ ) and the 480 ms condition ( $p < .001$ ). The total task durations were also found to be significantly shorter in the 380 ms condition than the 480 ms condition ( $p < 0.001$ ). All other comparisons were found to be non-significant ( $p > .05$ ).

#### 4.2. Error rate

The error rates were taken as the number of target locations which required one or more repeats per block (including the not-in-list option). Statistical analysis was performed using generalised estimating equations (GEE) [23]. GEE analysis was chosen because the observed error rates violate the assumptions of normality required for traditional ANOVA-based methods. As the dependent variable was count data, the model was constructed using a Poisson distribution and a log-link function. The working correlation matrix was specified as auto-regressive (AR(1)) because error rates were likely to be more correlated with neighbouring onset asyn-



(a) Task duration



(b) Error-rate

Figure 3: (a) Marginal means of the square root transformed total task durations (b) Marginal means (original scale) for the number of trials requiring one or more repeats during each block of 10 target locations. The 360 and 600 ms word durations are represented by the solid line with filled markers and the dotted line with hollow markers respectively. (Error bars =  $\pm 1 S.E.$ )

chrony/word duration conditions. Convergence criteria were set as an absolute difference between iterations of less than  $10^{-6}$ .

The model fit values were 118 and 122 (to 3 s.f.) for the quasi likelihood under independence model criterion (QIC) and the corrected quasi likelihood under independence model criterion (QICC) respectively. Results of the model indicated that the effects of onset asynchrony ( $Wald \chi^2(3) = 113, p < .001$ ), word length ( $Wald \chi^2(1) = 26.7, p < .001$ ) and their interaction ( $Wald \chi^2(3) = 37.0, p < .001$ ) were significant (Fig. 3b). *Post hoc* Bonferroni-corrected pairwise comparison of the interaction indicated that for the 360 ms stimuli all onset asynchronies were significantly different from each other with the exception of the 380 and 480 ms conditions. For the 600 ms word duration stimuli no adjacent onset asynchronies were found to provide significant improvements, although each condition was found to be significantly different from all others. Word durations were significantly different for the same asynchrony only in the 380 and 480 ms asynchrony conditions.

### 4.3. Workload

The unweighted scores produced by the TLX software [21] were used, which took the mean of the sub-scale scores for each participant to one decimal place. Shapiro-Wilk and Mauchly tests indicated that the normality and sphericity assumptions were met ( $p > .05$ ). Results from an rm-ANOVA (onset asynchrony  $\times$  word duration) indicated a significant main effect ( $F(3, 45) = 36.3, p < .001, h_p^2 = .708$ ) for onset asynchrony but no significant effect from word length ( $F(1, 15) = 3.43, p = .084, h_p^2 = .186$ ) or the interaction ( $F(3, 45) = .617, p = .608, h_p^2 = .040$ ). *Post-hoc* Bonferroni-corrected pairwise comparisons for onset asynchrony indicated that all of the treatments were significantly different from each other, with the exception of the 380 and 480 ms conditions.

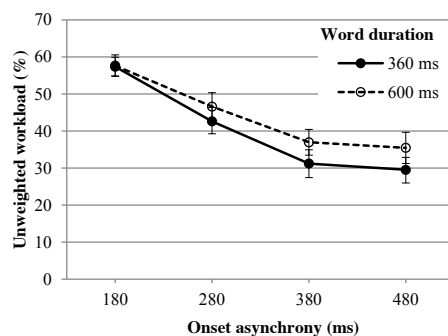


Figure 4: Marginal means for the unweighted TLX scores for the 360 (solid line with filled marker) and 600 ms (dashed line with hollow marker) word durations. (Error bars =  $\pm 1 S.E.$ )

## 5. DISCUSSION

### 5.1. Task duration

Due to the implicit effect of shortened words on the time taken to present information, it is of little surprise that the word duration factor exhibited a large significant effect on task durations. The *post hoc* analysis of the effect of the onset asynchrony on the total task duration indicates an optimum asynchrony of around 280 ms, despite this representing considerably different durations of overlap between the two word duration conditions.

The lack of a significant interaction between the word duration and onset asynchrony conditions suggests that the degree of asynchrony, as opposed to the proportion of overlapping stimuli, was most important in determining the time taken on each task. For the onset asynchrony conditions above 280 ms it is possible that reaction speed advantage is present due to the words being more easily identifiable. As the task durations increase, however, any improvement in time taken to detect the target is less than the increase in presentation time when the asynchrony is at its maximum of 480 ms.

### 5.2. Error rates

The observed interaction in the error rates appears to be due to diverging error rates for the two word duration conditions as the asynchrony increased, with the difference becoming significant at

the 380 and 480 ms asynchrony conditions. At these asynchronies the shorter stimuli are no longer overlapping, whereas the longer stimuli still overlap with the following word. This fact is particularly pertinent when it is recognised that the overlaps involve the endings of two words in each triplet, which in this task can be seen as the critical section for distinguishing between the maskers and the target word. This contrast in acoustic conditions was evidently more significant than between varying degrees of overlap in the smaller asynchrony conditions. It is, however, notable that the difference between the word durations in the 280 ms asynchrony conditions is considerable and it is speculated that an increased sample size might have led to significance.

The fall in error rate data over onset asynchrony appears generally in agreement with results from other studies on onset asynchrony [24, 12]. However, error rates appear to be higher than those found by [12] in equivalent conditions. Whilst it is possible that the use of intensity panning rather than binaural spatialisation may have contributed to the decreased accuracy, it seems unlikely that this difference alone would cause such a large discrepancy. It is also possible that modification of the word duration and pitch may have affected word intelligibility and inflated error values. However, it can be seen for both word durations that the accuracy approached 100% as the words became temporally distinct, implying that the processing of the words was not a major factor in itself. However, it is likely that the difference is predominantly due to the choice of experimental stimuli within this study. Whilst stimuli in this trial were distinguishable through only the final vowel-consonant transition, the words used by Ikei *et al.* [12] were more phonetically varied. This will have made the tasks considerably easier, as the increased phonetic variation will have provided the participants with more cues by which to distinguish the target word from the maskers.

It is possible that the increase in error rate observed here is responsible for an apparent disparity between the trend in error rate found by Ikei *et al.* [12] and the one found in this study. Whilst the results in [12] appeared to show an optimum onset asynchrony of 300 ms (for three voices and no attenuation), the results of this experiment appear to show further reduction in error rates up to the 480 ms condition for the longer stimuli. It is thought that this inconsistency is a by-product of the inflated error rate present in this study, and therefore the optimum asynchrony suggested in [12] is the product of a floor effect on error rates. Whilst [12] indicates that further accuracy could be achieved through the addition of ‘cross-ordering’ (presenting each word on the contralateral hemisphere to the preceding word) and applying an attenuation over the course of the word, neither of these methods were included in the design of the present study. Cross-ordering would not have been applicable due to the use of only three overlapping sources. It is feasible that through improving the audibility of word onsets, attenuation processing could have improved stream formation. In scenarios where the critical information is at the end of the word, however, the reduced signal-to-noise ratio is hard to justify.

Research into backwards recognition masking (BRM) indicates that vowel recognition performance plateaus when vowel onsets are separated by 200-250 ms or greater [25]. The range of asynchronies in the present study therefore suggests that BRM is unlikely to have been an influential factor for any asynchronies other than the 180 ms treatment. Due to the non-stationary nature of the speech signals used here, it could be that BRM impacted the stream formation and therefore made the location of the target more challenging to resolve.

It is notable that the constant location of critical information at the word ending may have led participants to listen only for the ending of the words and then use ordinal, spatial and, depending on the voicedness of the word ending, pitch information to derive which of the three locations the target had originated from.

### 5.3. Workload

The results of analysing the workload scores indicates that onset asynchrony was the only factor that influenced the participants' perception of task difficulty. In fact the workload scores appear to exhibit a divergent behaviour similar to error rate, though this difference was not large enough to be significant. Interestingly, this implies that the additional overlap associated with the longer stimuli did not significantly contribute to participants subjective workload in the two largest asynchronies, despite significantly increasing their error rate.

### 5.4. Overlap or onset asynchrony

It would appear that onset asynchrony describes observed trends for task duration and workload better than the amount of overlap. The error rate, however, displays a more complex interaction between the onset asynchrony and word length. The divergence between word durations with increasing asynchrony implies that both asynchrony and overlap are influential on performance. It is acknowledged that the difference between word durations was comparatively small due to the nature of the stimuli chosen and, therefore, based on this study it is not possible to come to any conclusion regarding situations in which the amount of overlap is considerably larger.

### 5.5. Asynchrony in menu display

Considering the effects of asynchrony on navigational speed, accuracy and subjective workload, it would appear that, of the treatments measured, the onset asynchrony of 380 ms provides the best compromise across all performance measures. Considering task durations alone, the lack of a significant difference between the 280 and 380 ms asynchrony conditions implies that an optimum exists between the two measured treatments. If one considers the additional time that would be incurred due to the higher error rates associated with the 280 ms condition, it seems likely that in practice this optimum is closer to the 380 ms condition. This conclusion is supported further through the workload scores, which show a significant reduction in workload from 280 to 380 ms onset asynchrony, suggesting that users felt that this condition made the interface significantly easier to use. The lack of overlap for this asynchrony condition for the shorter words, and its effect on error rate and navigational speed, is particularly pertinent, as it suggests that a more efficient solution would be to temporally compress the stimuli and present them with a short inter-stimuli interval.

Were a non-overlapping display to be used, a question is raised over whether the grouping of stimuli into triplets is advantageous. Grouping would seem likely to increase speed, as the number of physical interactions with the interface are reduced. Previous work comparing grouped and individual presentations of temporally distinct spoken items, however, indicates that participants are able to navigate to target locations faster when words are presented one at a time [26]. The grouped display in [26] imposed 200 ms inter-stimuli delays, whereas the present study, when using the shorter stimuli and the 380 and 480 ms asynchronies creates inter-stimuli

delays of 20 and 120 ms, respectively. This suggests that faster navigation may have been possible by reducing the size of the inter-stimuli delay with minimal impact on workload or error rate. Further investigation is recommended to ascertain the effect of grouped displays with lower inter-stimuli delays on performance to inform the future design of spoken auditory displays.

The methodology presented here primes the user with a visual representation of the target word and therefore simulates only a user with a very clear idea of the item which they are looking for. In such circumstances, a search-based navigation is likely to prove more efficient. The methodology also implies a selective attention task in which the user need only listen out for one word within the list and ignore all others. This is distinct from what is required in a browsing task where a user would be expected to have to listen to a set of possible selections before making a choice. The methodology in the present study was adopted to reduce response variation due to possible target identity confusion and therefore represents the ideal scenario in terms of target knowledge.

This paper has focused on the experimental investigation into the effects of onset asynchrony in spoken menus. Cognitive and perceptual theories that surround the use of concurrent or serial speech within user interfaces have not been discussed as they are beyond the scope of this paper.

It is worth noting that the stimuli used within this study were quite short, which may have restricted the degree of stream formation that could occur, causing critical information to be missed, or its location/order/pitch to be unresolved. With longer, less informationally dense content, as in [14], users may have been able to orientate their attention more effectively towards a desired stream of speech. However, it is at present still unclear whether this would offer a significant advantage in terms of both time saved and accuracy.

## 6. CONCLUSION

The problem of providing users with non-visual menus capable of facilitating browsing behaviour is a considerable design challenge for auditory display. Due to the limitations of non-speech methods regarding the representation of dynamic, novel content, it would appear that speech-based methods are most appropriate. This work has sought to explore the feasibility of using asynchronous, overlapping speech for menu representation and to determine what effect this has on the speed of navigation.

An experiment in which participants attempted to find a target word within a list of words was performed so that task duration, accuracy and subjective workload could be assessed for different onset asynchronies and word durations. The results of this experiment indicate that though some speed advantage may be present, it appears to be small and not significantly better than using shorter or temporally-compressed words with some grouping. This approach appears to have the added advantage of improving accuracy and perceived workload.

## 7. ACKNOWLEDGEMENTS

The authors would like to thank those who volunteered to participate in the experiment and the colleagues who have been recorded speaking the stimuli as part of this project. The authors would also like to thank the reviewers for their comments and suggestions.

## 8. REFERENCES

- [1] D. Elswiler, S. Mandl, and B. Kirkegaard Lunn, "Understanding casual-leisure information needs: a diary study in the context of television viewing," in *Proc. 3rd Symp. Inform. Interaction in Context (IiX)*, New Brunswick, NJ, Aug. 2010, pp. 25–34.
- [2] M. M. Blattner, D. Sumikawa, and R. Greenberg, "Earcons and icons: their structure and common design principles," *Human-Comp. Interaction*, vol. 4, no. 1, pp. 11–44, Mar. 1989.
- [3] W. Gaver, "Auditory icons: using sound in computer interfaces," *Human-Comp. Interaction*, vol. 2, no. 2, pp. 167–177, Jun. 1986.
- [4] B. N. Walker, J. Lindsay, A. Nance, Y. Nakano, D. K. Palladino, T. Dingler, and M. Jeon, "Spearcons (speech-based earcons) improve navigation performance in advanced auditory menus," *Human Factors: J. Human Factors and Ergonom. Soc.*, vol. 55, no. 1, pp. 157–182, Feb. 2013.
- [5] Y. Borodin, J. P. Bigham, G. Dausch, and I. V. Ramakrishnan, "More than meets the eye: a survey of screen-reader browsing strategies," in *Proc. 2010 Int. Cross-Disciplinary Conf. Web Accessibility (W4A)*, Raleigh, NC, Apr. 2010, Article 13.
- [6] A. S. Bregman, *Auditory Scene Analysis: The Perceptual Organization of Sound*. Cambridge, MA: MIT Press, 1994.
- [7] J. C. Webster and P. O. Thompson, "Responding to both of two overlapping messages," *J. Acoustical Soc. Am.*, vol. 26, no. 3, pp. 396–402, May 1954.
- [8] M. A. Ericson, D. S. Brungart, and B. D. Simpson, "Factors That Influence Intelligibility in Multitalker Speech Displays," *Int. J. Aviation Psychology*, vol. 14, no. 3, pp. 313–334, 2004.
- [9] W. T. Nelson, R. S. Bolia, M. A. Ericson, and R. L. McKinley, "Spatial audio displays for speech communications: a comparison of free field and virtual acoustic environments," *Proc. Human Factors and Ergonomics Society Annu. Meeting*, vol. 43, no. 22, pp. 1202–1205, Sep. 1999.
- [10] C. Frauenberger and T. Stockman, "Patterns in auditory menu design," in *Proc. 12th Int. Conf. Auditory Display (ICAD)*, T. Stockman, L. V. Nickerson, C. Frauenberger, A. D. N. Edwards, and D. Brock, Eds., London, UK, Jun. 2006, pp. 141–147.
- [11] C. Frauenberger, Personal correspondence, 2013.
- [12] Y. Ikei, H. Yamazaki, K. Hirota, and M. Hirose, "vCocktail: multiplexed-voice menu presentation method for wearable computers," in *Proc. IEEE Virtual Reality Conf.*, Alexandria, VA, Mar. 2006, pp. 183–190.
- [13] P. Parente, "Clique: perceptually based, task oriented auditory display for GUI applications," PhD Thesis, University of North Carolina, 2008.
- [14] J. Guerreiro and D. Gonçalves, "Text-to-speeches: evaluating the perception of concurrent speech by blind people," in *Proc. ASSETS'14*, Rochester, NY, Oct. 2014, pp. 169–176.
- [15] V. Shafiro and B. Gygi, "Perceiving the speech of multiple concurrent talkers in a combined divided and selective attention task," *J. Acoustical Soc. Am.*, vol. 122, no. 6, pp. EL229–EL235, Dec. 2007.
- [16] C. J. Darwin, D. S. Brungart, and B. D. Simpson, "Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers," *J. Acoustical Soc. Am.*, vol. 114, no. 5, pp. 2913–2922, Nov. 2003.
- [17] "Pd-extended." [Online]. Available: <http://puredata.info/downloads/pd-extended>
- [18] A. S. House, C. E. Williams, M. H. L. Hecker, and K. D. Kryter, "Articulation-testing methods: consonantal differentiation with a closed-response set," *J. Acoustical Soc. Am.*, vol. 37, no. 1, pp. 158–166, Jul. 1965.
- [19] P. Boersma and D. Weenink, "Praat: doing Phonetics by Computer." [Online]. Available: [www.praat.org](http://www.praat.org)
- [20] B. R. Glasberg and B. C. J. Moore, "Derivation of auditory filter shapes from notched-noise data," *Hearing Research*, vol. 47, no. 1-2, pp. 103–138, Aug. 1990.
- [21] A. Cao, K. K. Chintamani, A. K. Pandya, and R. D. Ellis, "NASA TLX: software for assessing subjective mental workload," *Behavior Research Methods*, vol. 41, no. 1, pp. 113–117, Feb. 2009.
- [22] IBM, "IBM Knowledge Center," Jan. [Online]. Available: [http://www-01.ibm.com/support/knowledgecenter/SSLVMB\\_21.0.0/com.ibm.spss.statistics\\_21.kc.doc/pv\\_welcome.html](http://www-01.ibm.com/support/knowledgecenter/SSLVMB_21.0.0/com.ibm.spss.statistics_21.kc.doc/pv_welcome.html)
- [23] S. L. Zeger and K.-Y. Liang, "Longitudinal data analysis for discrete and continuous outcomes," *Biometrics*, vol. 42, no. 1, pp. 121–130, Mar. 1986.
- [24] J. H. Lee and L. E. Humes, "Effect of fundamental-frequency and sentence-onset differences on speech-identification performance of young and older adults in a competing-talker background," *J. Acoustical Soc. Am.*, vol. 132, no. 3, pp. 1700–1717, Sep. 2012.
- [25] D. W. Massaro, "Perceptual units in speech recognition," *J. Experimental Psychology*, vol. 102, no. 2, pp. 199–208, Feb. 1974.
- [26] J. Sodnik, G. Jakus, and S. Tomažič, "Multiple spatial sounds in hierarchical menu navigation for visually impaired computer users," *Int. J. Human-Comp. Stud.*, vol. 69, no. 1-2, pp. 100–112, Jan. 2011.

## **A.2 Television and Additional Media Activity: A Taxonomy**

Available from: [https://figshare.com/articles/Television\\_and\\_additional\\_media\\_activity\\_A\\_taxonomy/3856164](https://figshare.com/articles/Television_and_additional_media_activity_A_taxonomy/3856164)

## Television and additional media activity: A taxonomy

Charlotte Hoare<sup>1</sup> and Alistair Hinde<sup>2</sup>  
c.m.hoare@bath.ac.uk      afh508@york.ac.uk

<sup>1</sup>Centre for Digital Entertainment, University of Bath

<sup>2</sup>Audio Lab, University of York

September 23, 2016

### Abstract

Television experiences are no longer restricted to a single audio-visual stream experienced on a single television set. Additional media activities that users engage with around a television programme are also an important experiential consideration. These activities, however, encompass a wide range of experiences and there is not a well defined set of terms in use that can adequately differentiate between them. This work considers factors that may be used to distinguish different types of experience and presents a taxonomy to provide a structured language for designers, practitioners, and researchers.

©2016 Charlotte Hoare, Alistair Hinde

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International. To view a copy of this license visit:

<http://creativecommons.org/licenses/by-nc/4.0/>.

## 1 Introduction

The definition of a ‘television experience’ is evolving due, in part, to the ever-increasing range of available consumer devices and the ubiquity of the Internet. One way it is evolving is through the supplementation of a television programme with additional media activity. This may comprise checking emails while watching the programme, using a dedicated play-along app, or catching up with the latest fan theories before the next episode. These use cases clearly reflect a wide range of different user motivations and as such require distinct consideration. For researchers, designers, and programme makers, an important prerequisite to this is having a language and terminology to adequately describe the different additional media activities that supplement television programmes.

Currently, researchers use a small set of terms to describe the diverse range of additional media activities that users engage in relative to particular television programmes. This can make the body of work in this emerging field hard to interpret, and can mean that the same terms are used to describe additional media activities that are actually very different in nature. This leaves the body of work open to the drawing of potentially erroneous generalised conclusions.

The terms used have included ‘second screen’ [17, 14, 9], ‘second screening’ [8, 7], ‘second screen experience’ [2, 18], ‘media multitasking’ [3, 5], ‘companion content’ [17, 4] and ‘companion experience’ [2, 13], amongst others. As stated, these terms do not necessarily refer to what is expected—and common sense conclusions on the distinctions between the terms are dangerous to draw, given that single terms have been applied to very different additional media activities. For example, Schirra et al. use the term ‘second screen’ to refer to the live-tweeting of a television programme [17], whereas Neate et al. use it to refer to a dedicated application designed and built to accompany a specific programme [14]. Whilst the use of this term makes sense for each example in its own context, each example is very different. In the interest of clarity, a well-defined set of terms would be helpful to categorise the range of additional media activities that users engage in to supplement television programmes.

There have been a handful of previous attempts made to define the meaning of particular terms. The ‘2nd Screen Society’, an industry body, has provided a ‘lexicon’ of terms and their definitions, although it is relatively unstructured [1]. Furthermore, some of the terms overlap and even appear to contradict one another. For example, the lexicon states that the definition of a ‘second screen’ is:

A companion experience in which a consumer engages in relevant content on a second device, such as a smart phone, tablet or laptop while watching



something on the “first screen” (typically a television but not limited to the living room). [1]

That the definition of a ‘companion experience’ is:

A second-device activity that is specially designed, by the creator of the first screen content (or service provider partner), to enhance the entertainment experience or viewing outcome. This extends to any experience provided by the TV industry that acts as a counterpart to your TV consumption, delivered on a second screen. [1]

And that ‘second screening’ is:

The broadest definition of second screen use, this covers any second-device activity undertaken while watching TV or a live event. While watching a TV program, viewers may be writing an email on a laptop, looking up sports results on a smartphone, or reading the news on a tablet: this is the 21st century version of reading the paper while watching TV. [1][sic]

It is clear that the terms ‘second screen’ and ‘second screening’ represent quite different behaviours and scopes according to the 2nd Screen Society definitions—though they sound very similar.

McGill et al. in their review paper of 2015 catalogue a wide range of terms in use [12]. They reference a 2012 report from Google on the ‘new multi-screen world’ [19], which uses the terms ‘multitasking’ and ‘complementary usage’. Google use both terms to refer to using one or more devices at the same time as watching television, with ‘multitasking’ referring to unrelated activities and ‘complementary usage’ referring to related activities. However, this particular use of the term ‘multitasking’ is somewhat misleading—clearly using one or more devices at the same time as watching television is a multitasking behaviour, regardless of whether or not the activities are related to the programme being watched. Indeed, elsewhere the term is used to refer to both related and unrelated activities [6].

Ofcom recently introduced the terms ‘media-meshing’ and ‘media-stacking’ to refer to the additional media activity of users while watching television [15], with ‘media-meshing’ meaning interacting with (or communicating about) content related to the television show, and ‘media-stacking’ meaning engaging with content unrelated to the show. These terms, whilst not yet widely adopted, are clearer than some of the other terms in use. The

phraseology of the terms has a clear relationship to the activities they represent and are plainly differentiable from each other. However, the terms do not entirely cover the kind of granularity needed to adequately describe the full range of additional media activities that users engage in relative to a particular television programme. For example, there is a considerable difference between reading an IMDB<sup>1</sup> webpage about a programme you are watching and interacting with a dedicated app designed to accompany it—though both would correctly fall under the term ‘media-meshing’. So whilst the terms ‘media-meshing’ and ‘media-stacking’ are certainly the most useful, it is necessary that more granular terms are introduced for accuracy.

In an attempt to improve on the current state of affairs, this article aims to develop a language and terminology for researchers and practitioners to effectively describe additional media activities that users engage in to supplement television programmes. While most of these media activities are understood in their own right, their relationship to television programmes is not yet well understood, necessitating a useful language and terminology. This article aims to develop such a taxonomy.

## 2 Examples of additional media activities

To demonstrate the diversity of additional media activities that a user may engage in relative to television programmes, this section presents several imagined ‘user journeys’.

*Juan is watching a historical drama (the BBC’s adaptation of ‘Wolf Hall’<sup>2</sup>) and has installed an app on his tablet that was custom-built for this particular series. He likes learning more about the historical events and artefacts he sees in the show—and this app delivers that information to him at appropriate moments throughout the programme for him to peruse. For example, when some characters display shock at the rise of Thomas Cromwell, the app immediately delivers some information about the unlikelihood of a lawyer rising so high in Tudor times. As well as delivering this information to Juan while he is watching the programme, the app also archives all the information so that when the show has finished, and while Juan is waiting for the next programme in the series, he can use the app to have another browse of all the content that has been delivered so far.*

*Stephanie is watching a current affairs programme, where several political figures are debating issues of the day. She tends to monitor the ‘Top Tweets’ on the hashtag provided at the start of the show on Twitter<sup>3</sup> while she’s watching, so she can see people’s reactions to*

---

<sup>1</sup><http://www.imdb.com/>

<sup>2</sup><http://www.bbc.co.uk/programmes/p02gfy02>

<sup>3</sup><https://twitter.com/>

*the debates.*

*Liam is watching a nature programme about the Arctic Ocean and navigates to the Wikipedia<sup>4</sup> page of beluga whales on his smartphone to get a bit more depth of information about how they have evolved.*

*Jess is a massive fan of ‘Sherlock’<sup>5</sup> and she’s been waiting all year for the new series to begin. In the week before the series premiere, she sees that the blog of fictional character, Dr John Watson, has been resurrected and so she reads through his new posts in anticipation of the programme.*

*Before she goes to bed, Shachi is catching up with the news on her TV, and dealing with a few emails on her laptop at the same time.*

### 3 Categorisation of additional media activities

Additional media activities relative to a particular television programme can be described by their content, when and how they are experienced, and the degree to which they have been orchestrated.

Before moving on to describe how additional media activities will be categorised, it is worth highlighting a couple of points. Firstly, it is important to clarify the way in which this taxonomy considers the term ‘experience’. The term ‘experience’ could be used as an encompassing term to refer to a user’s entire experience of a television programme, including both the programme itself and any additional media activities undertaken by the user. It could also be used to refer to each component part of the user’s entire experience of a television programme. For example, the viewing of the television programme could be considered to be an experience in and of itself, as could each individual additional media activity. As an example, consider the case where a user begins their experience by watching a particular television programme. They then simultaneously use their tablet to navigate to the Wikipedia article for a particular character in the television programme, before moving on to check their social media accounts. Finally, they decide to look up the IMDB page for the programme they are watching. Thus, the user’s experience has encompassed the experience of several different additional media activities and could continue to do so. As we are considering the categorisation of separate additional media activities on an atomic level, this document uses the term ‘experience’ to refer to the experience of each separate additional media activity—though it is acknowledged that the term can also be used as an

---

<sup>4</sup><https://en.wikipedia.org>

<sup>5</sup><http://www.bbc.co.uk/programmes/b018ttws>

encompassing term.

Secondly, it is important to note that additional media activity, within this context, refers to activity not required by the user to experience the main programme in its standalone form. It should be clear, therefore, that conventional use of accessibility services (e.g. subtitles and audio-description to supplement information a user may otherwise be unable to access in the programme) is beyond the scope of this classification. Within this document, the use of these services is considered as part of the main programme itself.

### 3.1 Relatedness

The first way that an additional media activity may be categorised is by considering the relatedness of the additional content being experienced. Indeed, existing terminology has attempted to capture this in the past, with related and unrelated activities featuring in some of the previously defined terminology presented in the introduction [19, 15].

In many cases the related nature of content may be clear to all (e.g. a webpage about the episode being watched). In other cases, however, the link may only be clear to the user (e.g. looking at information about a location that the user was reminded of as a result of a scene in the programme). Furthermore, additional content experienced both before and after the main programme may be considered related by the user. A user, for example, may hear a radio segment about a new television programme before deciding to watch it and then, after watching the show, read fan-site forums. It is, therefore, proposed that additional media activities are categorised based on the user's perception of the relatedness of the additional content at the time that they experience it.

There is one situation which may arise occasionally: the situation where a user may realise that content which they had consumed at an earlier date is related only during their experience of the main programme (e.g. 'this programme reminded me of the article I read last week'). For completeness, and given that additional media activities are categorised based on the user's perception of the relatedness of its content at the time that they experience it, this work categorises such cases as being unrelated.

### 3.2 Causality

Another way of categorising additional media activities is by considering the causality of the additional content: has the additional content been created or curated as a result of the television programme? Some content may have been created as a result of the main programme, while other content may have been produced irrespective of the main

programme's existence. Content produced as a result of the main programme may be from the persons involved within the creation of the main programme itself, from third party organisations, or from other members of the public. It may include promotional materials (such as adverts, official websites and apps), or unofficial websites and apps from third parties, or even social media posts from other users.

It is important to recognise that there are many scenarios in which the creator of accompanying media (e.g. a website or app) may borrow from, or reference, pre-existing resources—resources that were not created or curated as a result of the television programme. These resources may have no causal link with the main programme but be included as elements within an experience that was created specifically for the show. This curation is, in itself, an important distinction. Whether performed by human selection or algorithmically, this curated experience may be contrasted to non-curated experiences in which a user accesses resources which were not curated as a result of the main programme.

Curated experiences may be re-packaged to make it practically impossible for the user to distinguish information with no causal relationship with the main programme and content which was created as a direct result of the show. The taxonomy, therefore, does not distinguish between curation and creation, and divides content based on whether or not it was created/curated as a result of the show. Here, examples in which the content was created or curated due to the existence of the main programme are referred to as companion content. Within companion content, no distinction is made between automatically curated content and curation that is performed by a person. The perspective of the work is that by seeding the automatic curation system (e.g. creating a hashtag, specifying a search term or user-group to collect data from) a curatorial step has been taken. Furthermore, the process of curation is, in many cases, unlikely to be apparent or important to the end-user.

Companion content is contrasted with non-companion content, in which neither the curation nor the creation was due to the show. Within the taxonomy, the companion/non-companion distinction is only made for related content. It is theoretically possible that a user may not be aware that content is related to a show despite the content having been curated or created as a result of a show. This distinction is not useful, however, as from the user's perspective they are equivalent.

### 3.3 Synchronicity

Another factor at play is the question of when an additional media activity is undertaken by the user. If the user undertakes additional activity at the same time as they are watching a particular television programme, then the experience they are having is synchronous. If the activity is undertaken not at the same time, then the experience is asynchronous. This

factor has been somewhat captured in existing terminology, with terms like ‘simultaneous usage’ [19] or ‘multi-tasking’ [6] employed, but the asynchronous use case has gone largely unconsidered in the field.

Whilst this distinction is largely clear, there is some complexity that should be noted. Consider, for example, the case where a user pauses a programme during a synchronous experience to engage further with the additional content, or the case where a user continues with their additional activity after the programme has finished. It could be argued that in the first case, the experience as a whole is still a synchronous one—despite the user pausing the programme for a period—as the expectation is that the user will un-pause the programme when they have completed this activity. Indeed, in the second case, the same could be argued if the user is simply finishing off any activity they had undertaken. However, if the user ends one activity and then begins another—such as finishing one article on Wikipedia and then beginning another—it could be argued that the user has ended a synchronous experience and begun an asynchronous experience. This complexity in the distinction of an experience as either synchronous or asynchronous applies to edge cases only—as stated above, the distinction is usually clear. It is noted here in the interest of completeness and to recognise that a certain degree of common sense is required when classifying edge cases.

This stage of the categorisation, therefore, relates to the manner in which the user has chosen to experience the content: either synchronously or asynchronously. This taxonomy is primarily concerned with classifying those additional media activities that are undertaken by the user synchronously, as these activities are particularly variable and require further classification. It is, however, important to note the asynchronous use case.

### 3.4 Orchestration

Synchronous companion experiences may comprise applications that were built by someone other than the user with the express purpose of being experienced while watching the show (e.g. a play-along application). Alternatively, a synchronous companion experience could comprise general purpose resources that were created to be accessed in a wide range of scenarios (e.g. an IMDB page about the show). Within the taxonomy, these are referred to as orchestrated and improvised experiences, respectively.

The term orchestrated has, like many of the other terms used in this taxonomy, been used to describe experiences in the past. BBC R&D specified that ‘orchestrated media’ referred to the interaction, synchronisation, and collaboration of television programme and companion content across devices [11]. It is used in a similar fashion in this taxonomy to classify those synchronous companion experiences that exhibit such features. The term

improvised, though it has not been used in the field previously, is a useful counterpart to the term *orchestrated*—and can be used to effectively describe other, *unorchestrated* synchronous activities.

With *orchestrated* experiences, specific knowledge of the use case in the design stage allows for considerations of the user’s context within the show and, therefore, it is likely that the experience will be more tightly related to the context of the episode and even scene. This may also allow the creator to produce an experience that better complements the programme without risk of excessive distraction or spoilers. Furthermore, within an *improvised* experience, the user constructs their own experience of the show. This is a clear contrast to an *orchestrated* experience, in which the orchestrator has designed the experience for them.

Orchestration may take several different forms. For example, a programme may provide a ‘hashtag’ for users to engage with during the programme or tell participants to go to a specific web-poll during the course of the programme. Orchestration, however, does not have to be performed by the programme-maker or their associates. A user-generated hashtag created for discussion of an element within an ongoing show is still considered *orchestrated*. It is notable, therefore, that explicit calls-to-action are not a necessary requirement of an *orchestrated* experience. They are, however, a strong indication that orchestration has taken place.

*Orchestrated* experiences may vary in terms of the amount of control exerted by an orchestrator upon the coordination of additional content. The orchestrator may produce a ‘locked-down’ package of content to accompany an episode. In contrast, they may produce an experience that incorporates different content depending on when the programme is viewed. One example of this would be the orchestrator automatically aggregating content from an external source (e.g. UGC) that fulfils some defined criteria. An orchestrator may apply different degrees of control over the content, ranging from a basic search aggregation to a manual review and editorial. This factor, therefore, may be considered to be a scale with experiences in which the orchestrator exerts full control over content at one extreme, and experiences in which the orchestrator exerts minimal control at the other.

Orchestrators may also choose to control the time at which a user experiences additional content. Again, the orchestrator is faced with a scale of control. At one extreme an orchestrator may wish to take precise control over the time that content is presented to the user, so that it reaches them at an optimum moment. Conversely, an orchestrator may choose to hand over all control of the timing to the user. Again, intermediate points exist where the orchestrator may control the timing of collections of information that the user is able to navigate at their leisure.

It is, therefore, proposed to categorise media activities based on the presence of an or-

chestrator for the experience. By definition, orchestrated experiences consist of content that has either been created or curated for the show, and are therefore also companion experiences. This distinction of orchestration, however, is based on whether the way in which a user chooses to experience it was explicitly intended. To demonstrate this contrast, consider a website that has been built to accompany a series that is populated with content such as character profiles, behind the scenes footage etc. Synchronous use of the site would not be orchestrated as the site was not created with the express purpose of providing this experience. Conversely, if the site was to provide information specifically to be experienced during the show such as social media conversations about the ongoing programme, this is both companion content and orchestrated.

Within orchestrated content, a distinction is made regarding how fully defined the additional content is by the orchestrator. Though it has been recognised that this factor exists on a scale, two categories are defined. Time-invariant produced packages of content are referred to as fixed experiences, while all other instances are referred to as evolving. To illustrate this division, consider an application that provides the user with specially selected photos to look at during a show. If the content of the application would be the same if a user were to watch the same programme a year later, then the experience is fixed. If, however, the images are selected periodically from the most highly-rated on a site, this would be considered an evolving experience.

A further step of categorisation refers to the degree of control the orchestrator has taken over the timing of content within the experience. Again, while it is acknowledged that different levels of control may be taken by the orchestrator, three distinct categories are put forward. Where each element of content is delivered at specific points decided by the orchestrator this is referred to as scheduled. Where the orchestrator has not taken any control of timing (i.e. the user is in full control of when content is experienced), this is referred to as unscheduled. Experiences which fall between these extremes are considered partially-scheduled. An example of such an experience could be an experience where the content is delivered in ‘chapters’, one after each commercial break. While the timing of the chapter delivery is controlled by the orchestrator, the user is given control over when to access elements within each chapter.

### 3.5 Devices

One manner in which additional media activities may vary is in terms of the devices used and, by extension, the modalities that they exploit. For example, one may distinguish between additional media that is presented on the same screen as the device (e.g. Touch-



cast<sup>6</sup>) and those additional media activities that occur on an extra device (e.g. playing a game on a phone while watching the television).

It is, however, also important to think beyond screen-based activities. With the current interest in the development of connected objects, there is the potential for media experiences that involve Internet-of-Things (IoT) devices. Such experiences offer the potential for ever-more creative additional media activities that escape the confines of the screen (e.g. a toy that acts out action from the programme [10]). Furthermore, users may require or choose to undertake additional media activities in different modalities (e.g. audio or braille) either due to access needs or preference.

Though device and, by extension, modality are factors that could be considered within a taxonomy they are not included within this work for several reasons. Firstly, as an emerging area there is a lot of potential in terms of devices and modalities that have yet to be explored. It therefore seems premature to impose a structure upon them. Secondly, the removal of device and modality from the taxonomy is to demonstrate the importance of developing equivalent experiences for those with different access needs and device limitations. It is hoped that by doing this it will encourage practitioners to consider the design of such equivalence in future experiences.

It is recognised that not introducing device or modality into the taxonomy is not without issue. Devices and modalities clearly have different sets of limitations and, therefore, design considerations. Furthermore, experiential differences may be significant to the user. Content that may work well in one modality may require alterations to be suitable for another (e.g. made shorter, timed or ordered differently). It should, therefore, be clear that their omission from the taxonomy does not mean device and modality do not require reporting or consideration at a later point.

### 3.6 User activity

The amount that a user is actively involved in the experience may also vary considerably. One can envisage scheduled orchestrated experiences that simply display information to the user and require no interaction. Conversely, play-along applications may require considerable user interaction. There are also a host of intermediate conditions requiring some interaction. An interesting case is social media activity, which can encompass both extremes, with some users choosing to watch the conversation and others choosing to actively engage with it [16].

From this it is clear that user activity is a continuum. In a similar manner to devices and

---

<sup>6</sup><http://www.touchcast.com/>

modality, user activity is not considered as a categorisation step within this taxonomy, but highlighted as a modifying factor that is important to report.

## 4 Taxonomy structure

A structure is provided for the classification of additional media activities based on the factors introduced in the previous section (See Figure 1).

The first categorisation step distinguishes between *related* and *unrelated* content. Related content is then further divided into either *companion* or *non-companion* content. This step is omitted from the unrelated branch because, given the user considers the content to be unrelated, its curation or creation is inconsequential.

A categorisation is then based on whether they are experienced *synchronously* or *asynchronously*. This applies to all branches, as users may choose to access any of the content types at any time relative to the programme. Synchronous companion content is then categorised further based on orchestration, as either *orchestrated* or *improvised*. For non-companion content or unrelated content, however, the synchronicity is the final classification, as they are implicitly improvised. As this work is concerned primarily with synchronous use, further categorisation of asynchronous experiences is not recommended here.

Orchestrated experiences are then categorised further based on the amount of control taken by the orchestrator. First they are categorised as *fixed* or *evolving* and then how scheduled they are. They are then categorised as either *scheduled*, *partially-scheduled*, or *unscheduled*.

## 5 Categorising the examples of additional media activities

To demonstrate the taxonomy, the examples given earlier are categorised in this section.

Juan's experience of using a custom-built app while watching a historical drama would be categorised as an orchestrated, synchronous companion experience. It would further be described as fixed and scheduled. If Juan chooses to browse the app between programmes as described in the user journey, then that experience would be categorised as an asynchronous companion experience.

Stephanie's experience of using Twitter while watching a current affairs programme would also be considered to be an orchestrated, synchronous companion experience. This one,

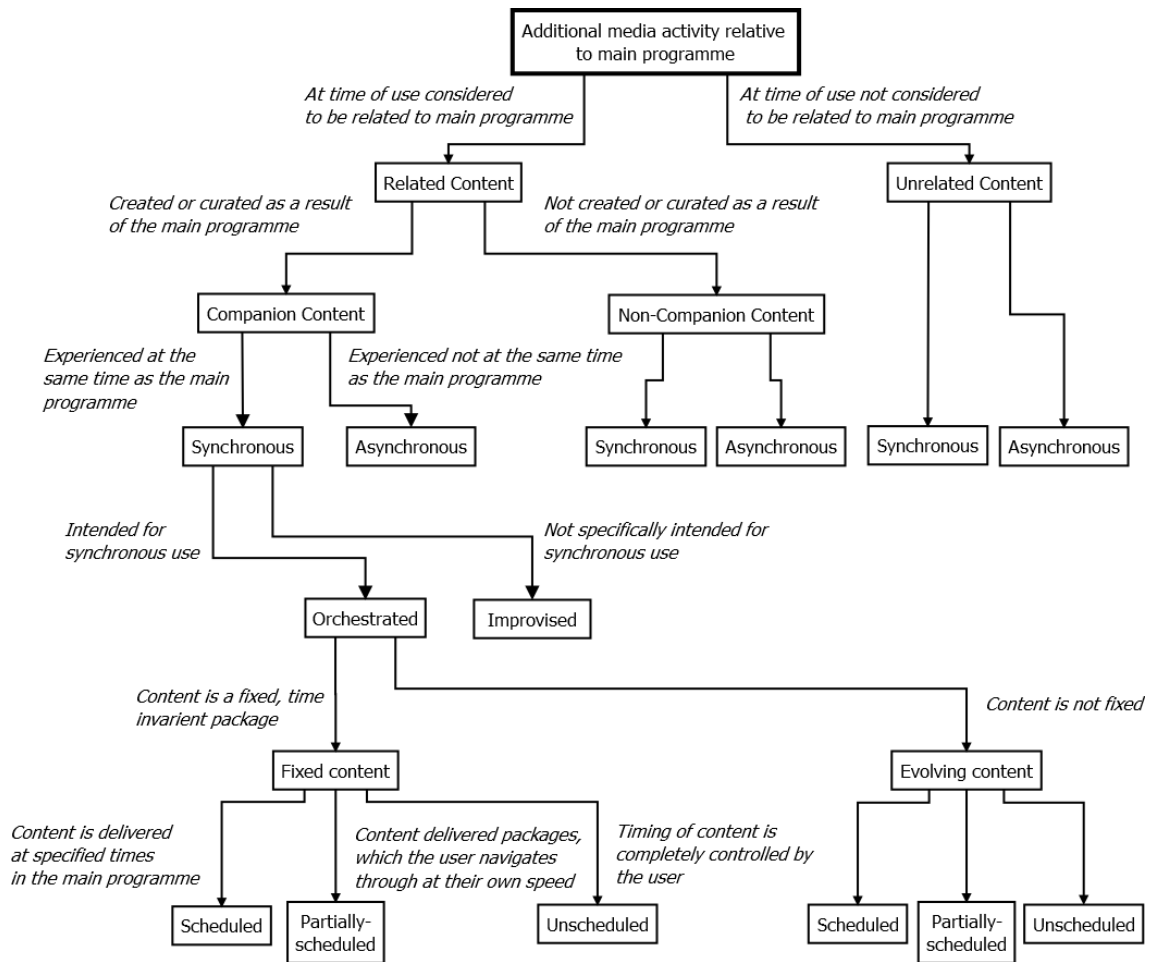


Figure 1: Graphical representation of the proposed taxonomy for additional media activity relative to a main programme

however, would be categorised as evolving and unscheduled. The fact that Stephanie is looking at the ‘Top Tweets’ on a particular hashtag shows that the content she is viewing has been both created (by users of Twitter) and curated (by the ‘Top Tweets’ algorithm) for the programme. As she is using Twitter whilst watching the programme, her experience is synchronous. The fact that she uses a defined hashtag that has been created for use during the show makes her experience orchestrated. However, as the orchestrator (i.e. whomever came up with the hashtag) has no control over what the people of Twitter will say on the hashtag, Stephanie’s experience is considered to be evolving. Equally, as they have no control over the timeliness of the delivery of the tweets, her experience is considered to be unscheduled. Thus, Stephanie’s experience is categorised as an orchestrated, synchronous companion experience that is evolving and unscheduled.

Liam’s experience of exploring more about beluga whales on Wikipedia while watching a nature programme is categorised as a synchronous, non-companion experience—as the Wikipedia article was not created or curated for the programme.

Jess’s experience of reading the blog of Dr John Watson, a fictional character from the drama ‘Sherlock’, before the premiere of the new series would be categorised as an asynchronous companion experience.

Shachi’s experience of checking her emails whilst watching the news would be categorised as an unrelated synchronous activity.

## 6 Conclusion

For researchers considering the supplementation of television programmes with additional media activities, the ill-defined and overloaded terminology that is currently in use can be problematic. This work has sought to introduce a taxonomy that can describe the full range of possible experiences, and is granular enough to differentiate those experiences from each other. The work has intentionally taken a device-agnostic perspective. A number of factors were identified that could be used to effectively delineate experiences from each other. These included their content’s relatedness to the television programme, whether their content was created/curated as a result of the television programme, the time at which they are experienced, and their degree of orchestration.

A taxonomy has been introduced, using these factors, to differentiate additional media activities within a structured set of terms. It is hoped that this taxonomy will provide benefit to both researchers, in providing a clear language by which to refer to their work and that of others, and practitioners, in thinking about the design of new user experiences. Equally, it should be noted that though this taxonomy has been presented as a way of

categorising the additional media activities that a user engages with relative to a particular television programme, it is also hoped the taxonomy could be useful for the categorisation of additional media activities relative to other forms of media—a radio programme, for example. Indeed, in a world of transmedia storytelling, it is hoped that any aspect of an experience could be considered as the main focal point, depending on what is the main focal point for the user. As this field continues to develop, it is inevitable that new categories of additional media activity will emerge requiring further extensions to this taxonomy. It is hoped that this work will provide greater clarity in this field going forward and a strong foundation on which future categories of experience may be added.

## 7 Acknowledgements

The authors would like to thank Mike Evans, Stephen Jolly, Vinoba Vinayagamoorthy, Matt Hammond, and Danaë Stanton Fraser for their comments on earlier versions of this manuscript, and the UK's EPSRC for funding their studentships.

## References

- [1] 2nd Screen Society. Lexicon for the 2nd Screen Society. <http://www.2ndscreensociety.com/lexicon/>.
- [2] Santosh Basapur, Hiren Mandalia, Shirley Chaysinh, Young Lee, Narayanan Venkataraman, and Crysta Metcalf. FANFEEDS: Evaluation of Socially Generated Information Feed on Second Screen As a TV Show Companion. In *Proceedings of the 10th European Conference on Interactive TV and Video*, EuroITV '12, pages 87–96, New York, NY, USA, 2012. ACM.
- [3] S. Adam Brasel and James Gips. Media Multitasking Behavior: Concurrent Television and Computer Usage. *Cyberpsychology, Behavior and Social Networking*, 14(9):527–534, September 2011.
- [4] Andy Brown, Michael Evans, Caroline Jay, Maxine Glancy, Rhianne Jones, and Simon Harper. HCI over Multiple Screens. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '14, pages 665–674, New York, NY, USA, 2014. ACM.
- [5] Duncan P. Brumby, Helena Du Toit, Harry J. Griffin, Ana Tajadura-Jiménez, and Anna L. Cox. Working with the Television on: An Investigation into Media Multitasking. In *Proceedings of the Extended Abstracts of the 32Nd Annual ACM Conference on Human Factors in Computing Systems*, CHI EA '14, pages 1807–1812, New York, NY, USA, 2014. ACM.

- [6] Ericsson ConsumerLab. TV and Media: Identifying the needs of tomorrow's video consumers. <https://www.ericsson.com/res/docs/2013/consumerlab/tv-and-media-consumerlab2013.pdf>, 2013.
- [7] Cédric Courtois and Evelien D'heer. Second Screen Applications and Tablet Users: Constellation, Awareness, Experience, and Interest. In *Proceedings of the 10th European Conference on Interactive TV and Video*, EuroiTV '12, pages 153–156, New York, NY, USA, 2012. ACM.
- [8] Mark Doughty, Duncan Rowland, and Shaun Lawson. Who is on Your Sofa?: TV Audience Communities and Second Screening Social Networks. In *Proceedings of the 10th European Conference on Interactive TV and Video*, EuroiTV '12, pages 79–86, New York, NY, USA, 2012. ACM.
- [9] David Geerts, Rinze Leenheer, Dirk De Grooff, Joost Negenman, and Susanne Heijstraten. In Front of and Behind the Second Screen: Viewer and Producer Perspectives on a Companion App. In *Proceedings of the 2014 ACM International Conference on Interactive Experiences for TV and Online Video*, TVX '14, pages 95–102, New York, NY, USA, 2014. ACM.
- [10] Stephen Jolly and Michael Evans. Improving the Experience of Media in the Connected Home with a New Approach to Inter-Device Communication. <http://www.bbc.co.uk/rd/publications/whitepaper242>, 2013.
- [11] Jerry Kramskoy. Orchestrated Media: beyond second and third screen (II). <http://www.bbc.co.uk/blogs/researchanddevelopment/2011/02/orchestrated-media---beyond-se.shtml>, 2011.
- [12] Mark McGill, John H. Williamson, and Stephen A. Brewster. A review of collocated multi-user TV. *Personal and Ubiquitous Computing*, 19(5-6):743–759, June 2015.
- [13] Abhishek Nandakumar and Janet Murray. Companion Apps for Long Arc TV Series: Supporting New Viewers in Complex Storyworlds with Tightly Synchronized Context-sensitive Annotations. In *Proceedings of the 2014 ACM International Conference on Interactive Experiences for TV and Online Video*, TVX '14, pages 3–10, New York, NY, USA, 2014. ACM.
- [14] Timothy Neate, Matt Jones, and Michael Evans. Mediating attention for second screen companion content. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 3103–3106, New York, NY, USA, 2015. ACM.
- [15] Ofcom. The Communications Market Report 2013. [http://stakeholders.ofcom.org.uk/binaries/research/cmr/cmr13/2013\\_UK\\_CMR.pdf](http://stakeholders.ofcom.org.uk/binaries/research/cmr/cmr13/2013_UK_CMR.pdf), 2013.

- [16] Margherita Pagani, Charles F. Hofacker, and Ronald E. Goldsmith. The influence of personality on active and passive use of social networking sites. *Psychology and Marketing*, 28(5):441–456, May 2011.
- [17] Steven Schirra, Huan Sun, and Frank Bentley. Together Alone: Motivations for Live-tweeting a Television Series. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems, CHI '14*, pages 2441–2450, New York, NY, USA, 2014. ACM.
- [18] Peter A. Torpey and Benjamin Bloomberg. Powers Live: A Global Interactive Opera Simulcast. In *Proceedings of the 11th Conference on Advances in Computer Entertainment Technology, ACE '14*, pages 16:1–16:9, New York, NY, USA, 2014. ACM.
- [19] Think with Google. The New Multi-Screen World Study. <https://www.thinkwithgoogle.com/research-studies/the-new-multi-screen-world-study.html>, 2012.





## Appendix B

# Menu navigation experiment documents

### B.1 Ethics Documentation

Ethical approval from was provided by the Physical Sciences Ethics Committee of the University of York for both studies discussed in Chapter 5. Approval for the pilot study was received on the 04/12/2013. The main experiment was given the reference number (hinde 140812) and approved on 20/10/2014.

### B.1.1 Pilot Ethics Documentation

PSEC/Application\_Form/2011.3



#### Application Form for Physical Sciences Ethics Committee approval

***Advice for applicants on completing the form***

*Please ensure that the information provided is:*

- *Accurate and concise*
- *Clear and simple and easily understood by a lay person*
- *Free of jargon, technical terms and abbreviations*

*Further advice and information can be obtained from your departmental representative on the PSEC and at: <http://www.york.ac.uk/admin/aso/ethics/ctee.htm>*

***Please return completed form to: Your departmental representative.***

Dr Martin Robinson, Department of Electronics

***Title of project:*** Non-visual display for connected television

#### **SECTION 1 DETAILS OF APPLICANTS**

**Details of Principal Investigator (name, appointment and qualifications)**

Alistair Hinde; PhD Student with the Audio Lab Research Group, but based in BBC R&D User Experience and Accessibility Group; MSc, BSc(Hons)

**Names, appointments and qualifications of additional investigators**

Anthony Tew; Senior lecturer; BSc, CEng, MIET, MIEEEE, AMAES, MIPEM

Prof. David Howard; Head of Department; BSc (Eng), PhD, CEng, FIET, FIOA, MAES

Mike Evans; Research lead in User Experience and Accessibility, BBC R&D; MEng, CEng, DPhil

**Location(s) of project**

BBC R&D (MediaCityUK, Salford) & University of York
---

**SECTION 2 FUNDERS****What is the funding source(s) for the project?**

ESPRC iCase studentship with University of York and BBC R&D
---

**Do the following principles apply?:**

- (i) The express and direct aim of the research is ethically defensible;

YES	X	NO	
-----	---	----	--

- (ii) There is no obvious or inevitable adaptation of research findings to ethically questionable aims;

YES	X	NO	
-----	---	----	--

- (iii) The work is not being funded by organisations tainted by ethically questionable activities;

YES	X	NO	
-----	---	----	--

- (iv) Restrictions on academic freedoms - notably, to adapt and withdraw from ongoing research, and to publish findings – are justifiable and minimal.

YES	X	NO	
-----	---	----	--

If **No** to any of the above, please give details below:

**SECTION 3 DETAILS OF PROJECT****Aims (100 words max)**

The aim of the project is to assess factors relating to non-visual display for connected television to improve user experience for those unable to see the screen. This experiment will investigate a range of scenarios for vocalising a list of words using speech. Specifically, it will present the words concurrently with different degrees of onset asynchrony and investigate the effects on navigational speed and workload.

**Background (250 words max)**

The navigation of menus and lists is a problem encountered often with television services. Simple serial presentation of a spoken wordlist can make the navigation of longer lists slow. Through presenting several concurrent streams of speech it is thought that it may be possible to achieve faster navigation of these structures.

Previous reported work has shown limited improvements to navigation speed from concurrent presentation (Frauenberger & Stockman, 2006).; however it is felt that this may have been due to the design of the displays used in the study. Ikei *et al.* (2006) proposed an interface using onset asynchrony and binaural separation to achieve high accuracy for navigation through auditory menus. However, no results were presented for the effect on navigational speed. To the author's knowledge no researchers have investigated the effect of differing degrees of onset asynchrony on navigational speed within spoken lists.

**Brief outline of project (250 words max)**

The purpose of this experiment is to investigate the effect of onset-asynchrony on navigation time and cognitive load when selecting a spoken cue from an auditory list. The system will split the list into groups of three spoken words (triplets) which will be presented with varying degrees of temporal overlap. A control condition will not group the items and instead present each item one at a time in isolation.

Users will be asked to navigate a list to find and select a target word as quickly as possible using drum-pads on a MIDI controller. The target word will be presented to them visually on-screen throughout each trial. The items will be represented as spoken cues, all spoken by the same talker with an artificially introduced fixed pitch difference. The experiment will be performed with three concurrent speech streams, one which will appear to come from the extreme left of the user, one from the middle and one from the right. The stimuli will be presented over headphones.

**Study design** (if relevant – e.g. randomised control trial; laboratory-based)

The controlled variables within the experiment are onset asynchrony (5 + 1 (serial presentation)) and location of target in the list (9). The identity of the target will be pseudo-random to avoid the same target in two consecutive trials (8 possible words) and the list contents other than the target will be pseudo-randomised, such that the target word cannot appear in any other location and the same word cannot be repeated in the same triplet. The experiment will record: a time-stamped log of user interactions, the ratings from a computer-based version of the un-weighted NASA Task Load Index (TLX) questionnaire (Cao et al., 2009), the audio output played to the user, and an audio recording of the sessions.

Prior to the experiment the user will be familiarised with the interface and given a few tasks similar to those of the experiment proper. For each of the onset asynchronies they will be asked to complete computer-based unweighted NASA TLX evaluations. This should take between 5 and 10 minutes. The results from this portion of the session will not be used in later analysis.

Each participant will perform all conditions leading to a total of 54 trials, which will be blocked by onset asynchrony and the order of these blocks will be varied to reduce the effect of any learning between conditions. Within each block the order of the target locations will also be randomised. The participants' interactions with the device will be time-stamped and logged in a text file. After each block the user will evaluate the workload for the display using a computer-based version of the un-weighted NASA TLX questionnaire. After all of the experiment blocks have been completed, participants will be asked a number of questions regarding their opinions of the interfaces, their strategies in the experiments and if they felt able to understand the non-target words (see attached script). The experiment will be split into two sessions the first of which will comprise of the introduction to the experiment and the training whilst the second will consist of the experiment proper and debriefing. Each session will be no longer than 20 minutes and a 20 minute break will be provided between the two sessions.

**If the study involves participants, how many will be recruited?**

8

**What is the statistical power of the study?**

This study will be a pilot and therefore the statistical power is not clear at this time. As the study is a pilot the number of participants and the duration of the experiment have been kept short.

**SECTION 4 RECRUITMENT OF PARTICIPANTS**

**How will the participants be recruited?**

For this study people within the BBC R&D department in MediaCityUK will be emailed and asked to volunteer to take part in the experiment.

**What are the inclusion/exclusion criteria?**

Participants will be from within BBC R&D.

The person whose voice is used in the experiment will not be a participant in the experiment, neither will supervisors or people who have been consulted on the design of the experiment. As the recruitment is within the office where the prototype was developed, it is likely participants will have some idea of the experiment, possibly heard the stimuli previously and in some cases used an earlier version of the prototype. Participants will not be excluded due to previous exposure to some of the stimulus or earlier prototypes, as it is felt to be unlikely to have a significant effect. However, they will be asked in the experiment whether they have previously heard the stimuli and if so to what extent, so that it may be taken into account in analysis of the results.

Only participants who classify their hearing as normal will be included in the experiment. Though no audiometric testing will be performed, the recruitment email will specify that participants with normal hearing are required.

**Will participants be paid reimbursement of expenses?**

YES	<input type="checkbox"/>
-----	--------------------------

NO	<input checked="" type="checkbox"/>
----	-------------------------------------

**Will participants be paid?**

YES	<input type="checkbox"/>
-----	--------------------------

NO	<input checked="" type="checkbox"/>
----	-------------------------------------

*If yes, please obtain signed agreement*

**Will any of the participants be students?**

YES	<input checked="" type="checkbox"/>
	(possibly)

NO	<input type="checkbox"/>
----	--------------------------

**SECTION 5 DATA STORAGE AND TRANSMISSION**

**If the research will involve storing personal data, including sensitive data, on any of the following please indicate so and provide further details (answers only required if *personal data* is to be stored).**

<b>Manual files</b>	Consent forms - kept in locked drawer at the BBC
<b>University computers</b>	
<b>Home or other personal computers</b>	An encrypted BBC computer (possibly a laptop) will be used in addition to Alistair's BBC laptop to run either the prototype or the NASA TLX software during the experiment; the recorded data will then be copied onto Alistair's BBC laptop and USB drive and deleted from the other computers used in the experiment.  The data files from the NASA TLX program and files containing data taken from the prototype's data logs. On BBC computer for analysis using SPSS
<b>Laptop computers</b>	Audio recordings of session (and transcriptions in word files) and system output, data log from prototype, data files containing relevant data extracted from log formatted for SPSS, data files from NASA TLX program. Stored on Alistair's encrypted BBC laptop*
<b>Website</b>	

\* An encrypted USB drive will also be used to transfer data between machines and serve as a backup. So as to avoid data loss, this key will be kept in the draw with the manual files when not in use. Also, the sessions will be initially recorded on a dictaphone but deleted when transferred onto the laptop.

**Please explain the measures in place to ensure data confidentiality, including whether encryption or other methods of anonymisation will be used.**

All data will be stored on an encrypted laptop and will be recorded in anonymised form using participant numbers only. Recordings from the dictaphone will be transferred onto the encrypted computer and the original files on the dictaphones will be deleted. The consent forms will be marked with participant number in pencil (so that a participant's data may be withdrawn from the trial on their request) and kept securely in a locked drawer at the BBC.

**Please detail who will have access to the data generated by the study.**

The persons noted as investigators

**Please detail who will have control of and act as custodian for, data generated by the study.**

Alistair Hinde

**Please explain where, and by whom, data will be analysed.**

Data will be analysed at BBC R&D in MediaCityUK by Alistair Hinde, with possible input from the other investigators and members of the User Experience and Accessibility Research Group at the BBC.

**Please give details of data storage arrangements, including where data will be stored, how long for, and in what form.**

Data will be stored as anonymised text, audio and SPSS files on Alistair’s laptop and on an encrypted USB drive or as paper copies (in the case of the consent form) in a locked draw for the duration of the PhD program, scheduled to be completed by April 2016.

**SECTION 6 CONSENT**

**Is written consent to be obtained?**

YES	<input checked="" type="checkbox"/>
-----	-------------------------------------

NO	<input type="checkbox"/>
----	--------------------------

*If yes, please attach a copy of the information for participants*

*If no, please justify*

**Will any of the participants be from one of the following vulnerable groups?**

Children under 18	YES	<input type="checkbox"/>	NO	<input checked="" type="checkbox"/>
People with learning difficulties	YES	<input type="checkbox"/>	NO	<input checked="" type="checkbox"/>
People who are unconscious or severely ill	YES	<input type="checkbox"/>	NO	<input checked="" type="checkbox"/>
People with mental illness	YES	<input type="checkbox"/>	NO	<input checked="" type="checkbox"/>
NHS patients	YES	<input type="checkbox"/>	NO	<input checked="" type="checkbox"/>
Other vulnerable groups	YES	<input type="checkbox"/>	NO	<input checked="" type="checkbox"/>

**If so, what special arrangements have been made for getting consent?**



**SECTION 7 DETAILS OF INTERVENTIONS**

**Indicate whether the study involves procedures which:**

Involve taking bodily samples

YES		NO	X
-----	--	----	---

Are physically invasive

YES		NO	X
-----	--	----	---

Are designed to be challenging/disturbing (physically or psychologically)

YES	X	NO	
-----	---	----	--

**If so, please list those procedures to which participants will be exposed:**

Participants will be asked to navigate as fast as possible and some of the display configurations may be challenging, therefore exposing participants to higher cognitive loads and possible stress.

**List any potential hazards:**

- (i) Hearing damage from listening using headphones. The possibility of a participant receiving hearing damage from the experiment will be minimised by using either sound pressure limited headphones or introducing an in-line limiter to keep the maximum level at a safe value. Also, the participants will be required to adjust the playback level (starting at the silent position) to one they find comfortable for listening. The duration of each session is limited so as to avoid excessive noise exposure.
- (ii) There is also a risk of electrocution from the use of electronics. The possibility of any harm coming to the participant will be minimised by ensuring all devices running from mains power have been PAT tested.
- (iii) There is a risk of tripping due to the requirement of cables. However, the researcher will minimise this risk by routing cables so that they are out of the way where possible and taping down any loose cables.

**List any discomfort or distress:**

Possible hearing discomfort from headphones

**What steps will be taken to safeguard**

- (i) the confidentiality of information

All data will be recorded in anonymised form and stored on an encrypted system.

- (ii) the specimens themselves?

No specimens will be taken

**What particular ethical problems or considerations are raised by the proposed study?**

The users will be made aware that it is not they who are being tested but the device and that they are free to remove themselves from the experiment at any time.

The words used in the study have been chosen so as to avoid probable emotive subjects.

**What do you anticipate will be the output from the study? Tick those that apply:**

- Peer-reviewed publications
- Non-peer-reviewed publications
- Reports for sponsor
- Confidential reports
- Presentation at meetings
- Press releases

**Is there a secrecy clause to the research?**

YES

NO

**SECTION 8 SIGNATURES**

The information in this form is accurate to best of my knowledge and belief and I take full responsibility for it.

I agree to advise of any adverse or unexpected events that may occur during this project, to seek approval for any significant protocol amendments and to provide interim and final reports. I also agree to advise the Ethics Committee if the study is withdrawn or not completed.

Signature of Investigator(s):

*[Signature]* (14/2/11/2013)  
 .....  
*[Signature]* (14/11/2013)  
 .....

Date:

.....

- Responsibilities of the Principal Researcher following approval**
- If changes to procedures are proposed, please notify the Ethics Committee
  - Report promptly any adverse events involving risk to participants

## B.1.2 Main experiment Ethics Documentation

PSEC Application Form V3

THE UNIVERSITY *of York*

### Application Form for Physical Sciences Ethics Committee Approval

***Advice for applicants on completing the form***

*Please ensure that the information provided is:*

- *Accurate and concise*
- *Clear and simple and easily understood by a lay person*
- *Free of jargon, technical terms and abbreviations*

*Further advice and information can be obtained from your departmental representative on the PSEC and at: <http://www.york.ac.uk/admin/aso/ethics/cttee.htm>*

***Please return completed form to your departmental representative:***

Dr Helena Daffern, Department of Electronics

***Title of project: Non-visual Display for Connected Television***

#### **SECTION 1 DETAILS OF APPLICANTS**

**Details of principal investigator (name, appointment and qualifications)**

Alistair Hinde; PhD Student with the Audio Lab Research Group, but based in BBC R&D User Experience and Accessibility Group; MSc, BSc(Hons)

**Names, appointments and qualifications of additional investigators** (*student applicants should include their project supervisor(s) here*)

Anthony Tew; Senior lecturer; BSc, CEng, MIET, MIEEE, AMAES, MIPEM

Prof. David Howard; Head of Department; BSc (Eng), PhD, CEng, FIET, FIOA, MAES

Mike Evans; Research lead in User Experience and Accessibility, BBC R&D; MEng, CEng, DPhil

19<sup>th</sup> March 2014

**Location(s) of project**

BBC R&D (MediaCityUK, Salford) & University of York

**SECTION 2 FUNDERS**

**What is the funding source(s) for the project?**

ESPRC iCase studentship with University of York and BBC R&D

**Please answer the following:**

(i) Does the express and direct aim of the research or other activity raise ethical issues?

YES  NO

(ii) Is there any obvious or inevitable adaptation of research findings to ethically questionable aims?

YES  NO

(iii) Is the work being funded by organisations tainted by ethically questionable activities?

YES  NO

(iv) Are there any restrictions on academic freedoms – notably, to adapt and withdraw from ongoing research, and to publish findings?

YES  NO

If you answered **Yes** to any of the above, please give details below:

**SECTION 3 DETAILS OF PROJECT OR OTHER ACTIVITY****Aims (100 words max)**

The aim of the project is to assess factors relating to non-visual display for connected television to improve user experience for those unable to see the screen. This experiment will investigate a range of scenarios for vocalising a list of words using speech. Specifically, it will present the words concurrently with different degrees of onset asynchrony and investigate the effects on speed or task duration, accuracy, workload and participants' subjective opinions.

**Background (250 words max)**

The navigation of menus and lists is a problem encountered often with television services. Simple serial presentation of a spoken wordlist can make the navigation of longer lists slow. Through presenting several concurrent streams of speech it is thought that it may be possible to achieve faster navigation of these structures.

Previous reported work has shown limited improvements to navigation speed from concurrent presentation (Frauenberger & Stockman, 2006).; however it is felt that this may have been due to the design of the displays used in the study. Ikei *et al.* (2006) proposed an interface using onset asynchrony and binaural separation to achieve high accuracy for navigation through auditory menus. However, no results were presented for the effect on navigational speed. To the author's knowledge no researchers have investigated the effect of differing degrees of onset asynchrony on navigational speed or task durations within spoken lists. Prior to this experiment, an experiment was performed but technical issues and a methodological oversight meant it was not possible to statistically prove or disprove the hypothesis. This experiment's method attempts to address these failings and simplify the analysis. This experiment also includes a two word lengths in an attempt to separate the effect of onset asynchrony from the proportion of words overlapping.

**Brief outline of project/activity (250 words max)**

The purpose of this experiment is to investigate the effect of onset-asynchrony on the speed/duration and accuracy of navigation tasks, workload and subjective opinion when selecting a spoken cue from an auditory list. The system will split the list into groups of three spoken words (triplets) which will be presented with varying degrees of onset asynchrony.

Users will be asked to navigate a list to find and select a target word as quickly as possible using drum-pads on a MIDI controller. The target word will be presented to them visually on-screen throughout each trial. The items will be represented as spoken cues, all spoken by the same talker with an artificially introduced fixed pitch difference. The experiment will be performed with three concurrent speech streams, one which will appear to come from the extreme left of the user, one from the middle and one from the right. The stimuli will be presented over headphones.

**Study design** (if relevant – e.g. randomised control trial; laboratory-based)

The experiment will be a within-participants design, with each participant completing trials for all target locations, word length and onset asynchrony combinations. In order to gather data for correct navigations for all factor combinations, any incorrect trials or trials that are judged by the experimenter to have been affected by external influences will be repeated. In these cases, the word lists will be varied to be different to the last instance of the trial. When repeated trials are required additional trials will be added to ensure suitable gaps are present between the original and repeated trial. When more than one accurate trial is present for a given combination of word-length and onset asynchrony only the first will be used in the navigation time or speed analysis.

The experiment will record: a time-stamped log of user interactions, the ratings from a computer-based version of the un-weighted NASA Task Load Index (TLX) questionnaire (Cao et al., 2009), the audio output played to the user, and an audio recording of the sessions.

Trials will be blocked by onset asynchrony and word length. Each block will contain each target location at least once, and at least one trial in which the target is not present in the list. The order of the blocks will be varied to reduce the effect of any learning between conditions. After each block the user will evaluate the workload for the display using a computer-based version of the un-weighted NASA TLX questionnaire. After all of the experiment blocks have been completed, participants will be asked a number of questions regarding their opinions on the displays they heard, the effect of word lengths, their strategies in the experiments and if they felt able to understand the non-target words.

The experiment will be split into three sessions. In the first session the user will be familiarised with the interface with an instruction sheet and given a few tasks similar to those of the experiment proper and the opportunity to ask any questions. For each of the onset asynchronies they will be asked to complete computer-based unweighted NASA TLX evaluations. The results from this portion of the session will not be used in later analysis.

After the training session there will be a 20 minute break before the first of the experimental trials. The listening test portions of the experiments will be limited to 20 minutes each with 20 minutes between them.

**If the study involves participants, how many will be recruited?**

16

**If applicable, what is the statistical power of the study, i.e. what is the justification for the number of participants needed?**

Due to the changes that have been made to the tasks since the last experiment the degree of variance that can be expected for the conditions and therefore statistical power is not clear. To allow counterbalancing of the word length and onset asynchrony a minimum of 8 participants would be required. Twice this number has been chosen to increase the power of the test as the data is still expected to be noisy.

**SECTION 4 RECRUITMENT OF PARTICIPANTS****How will the participants be recruited?**

Staff at the BBC recruited by email

**What are the inclusion/exclusion criteria?**

Participants will be recruited from within the BBC but outside of the R&D department, to avoid excessive pre-existing knowledge of the work affecting the results.

Only participants who classify their hearing as normal will be included in the experiment. Though no audiometric testing will be performed, the recruitment email will specify that participants with normal hearing are required.

**Will participants be paid reimbursement of expenses?**

YES

NO

**Will participants be paid?**

YES

NO

*If yes, please obtain signed agreement*

**Will any of the participants be students?**

YES

NO

**SECTION 5 DATA STORAGE AND TRANSMISSION**

**If the research will involve storing personal data, including sensitive data, on any of the following please indicate so and provide further details (answers only required if *personal* data is to be stored).**

<b>Manual files</b>	Consent forms - kept in locked drawer at the BBC
<b>University computers</b>	
<b>Home or other personal computers</b>	<p>An encrypted BBC computer will be used to run the prototype during the experiment; the recorded data will then be copied onto Alistair's BBC laptop and USB drive and deleted from the other computers used in the experiment.</p> <p>The data files from the NASA TLX program and files containing data taken from the prototype's data logs. On BBC computer for analysis using SPSS</p>
<b>Laptop computers, tablets</b>	<p>The NASA TLX program will be run on Alistair's BBC laptop.</p> <p>Audio recordings of session (and transcriptions in word files) and system output, data log from prototype, data files containing relevant data extracted from log formatted for SPSS, data files from NASA TLX program. Stored on Alistair's encrypted BBC laptop*</p>
<b>Website</b>	

\* An encrypted USB drive will also be used to transfer data between machines and serve as a backup. So as to avoid data loss, this key will be kept in the draw with the manual files when not in use. Also, the sessions will be initially recorded on a dictaphone but deleted when transferred onto the laptop.

**Please explain the measures in place to ensure data confidentiality, including whether encryption or other methods of anonymisation will be used.**

All data will be stored on an encrypted laptop and will be recorded in anonymised form using participant numbers only. Recordings from the dictaphone will be transferred onto the encrypted computer and the original files on the dictaphones will be deleted. The consent forms will be marked with participant number in pencil (so that a participant's data may be withdrawn from the trial on their request) and kept securely in a locked drawer at the BBC.

**Please detail who will have access to the data generated by the study.**



PSEC Application Form V3

The persons noted as investigators

**Please detail who will have control of and act as custodian for, data generated by the study.**

Alistair Hinde

**Please explain where, and by whom, data will be analysed.**

Data will be analysed at BBC R&D in MediaCityUK by Alistair Hinde, with possible input from the other investigators and members of the User Experience and Accessibility Research Group at the BBC.

**Please give details of data storage arrangements, including where data will be stored, how long for, and in what form.**

Data will be stored as anonymised text, audio and SPSS files on Alistair's laptop and on an encrypted USB drive or as paper copies (in the case of the consent form) in a locked draw for the duration of the PhD program, scheduled to be completed by April 2016.

**SECTION 6 CONSENT**

**Is written consent to be obtained?**

YES	<input checked="" type="checkbox"/>	NO	<input type="checkbox"/>
-----	-------------------------------------	----	--------------------------

*If yes, please attach a copy of the information for participants*

*If no, please justify*

**Will any of the participants be from one of the following vulnerable groups?**

Children under 18	YES	<input type="checkbox"/>	NO	<input checked="" type="checkbox"/>
People with learning difficulties	YES	<input type="checkbox"/>	NO	<input checked="" type="checkbox"/>
People who are unconscious or severely ill	YES	<input type="checkbox"/>	NO	<input checked="" type="checkbox"/>
People with mental illness	YES	<input type="checkbox"/>	NO	<input checked="" type="checkbox"/>
NHS patients	YES	<input type="checkbox"/>	NO	<input checked="" type="checkbox"/>
Other vulnerable groups (if 'yes', please give details)	YES	<input type="checkbox"/>	NO	<input checked="" type="checkbox"/>

**If so, what special arrangements have been made for getting consent?**

**SECTION 7 DETAILS OF INTERVENTIONS**

**Indicate whether the study involves procedures which:**

Involve taking bodily samples	YES	<input type="checkbox"/>	NO	<input checked="" type="checkbox"/>
Are physically invasive	YES	<input type="checkbox"/>	NO	<input checked="" type="checkbox"/>
Are designed to be challenging/disturbing (physically or psychologically)	YES	<input checked="" type="checkbox"/>	NO	<input type="checkbox"/>

**If so, please list those procedures to which participants will be exposed:**

Participants will be asked to navigate as fast as possible and some of the display configurations may be challenging, therefore exposing participants to higher cognitive loads and possible stress.

PSEC Application Form V3

**List any potential hazards:**

- (i) Hearing damage from listening using headphones. The possibility of a participant receiving hearing damage from the experiment will be minimised by using either limited headphones or introducing an in-line limiter to keep the maximum level at a safe value. Also, the participants will be required to adjust the playback level (starting at the silent position) to one they find comfortable for listening.
- (ii) There is also a risk of electrocution from the use of electronics. The possibility of any harm coming to the participant will be minimised by ensuring all devices running from mains power have been PAT tested.
- (iii) There is a risk of tripping due to the requirement of cables. However, the researcher will minimise this risk by routing cables so that they are out of the way where possible and taping down any loose cables.

**List any discomfort or distress:**

Possible hearing discomfort from headphones

**What steps will be taken to safeguard**

- (i) the confidentiality of information

All data will be recorded in anonymised form and stored on an encrypted system.

- (ii) the specimens themselves?

No specimens will be taken

**What particular ethical problems or considerations are raised by the proposed study?**

The users will be made aware that it is not they who are being tested but the device and that they are free to remove themselves from the experiment at any time.

**What do you anticipate will be the output from the study? Tick those that apply:**

Peer-reviewed publications	<input checked="" type="checkbox"/>
Non-peer-reviewed publications	<input checked="" type="checkbox"/>
Reports for sponsor	<input checked="" type="checkbox"/>
Confidential reports	<input checked="" type="checkbox"/>
Presentation at meetings	<input checked="" type="checkbox"/>
Press releases	<input checked="" type="checkbox"/>

19<sup>th</sup> March 2014

PSEC Application Form V3

**Is there a secrecy clause to the research?**  
*If yes, please give details below*

YES

NO

**SECTION 8 SIGNATURES**

The information in this form is accurate to best of my knowledge and belief and I take full responsibility for it.

I agree to advise of any adverse or unexpected events that may occur during this project, to seek approval for any significant protocol amendments and to provide interim and final reports. I also agree to advise the Ethics Committee if the study is withdrawn or not completed.

Signature of Investigator(s):

*[Handwritten Signature]* 05/08/2014  
*[Handwritten Signature]* 05/08/2014

Date:

.....

***Responsibilities of the Principal Researcher following approval***

- If changes to procedures are proposed, please notify the Ethics Committee
- Report promptly any adverse events involving risk to participants

## B.2 Script

**When participant arrives:** Experimenter: Hi, I am [experimenter name], I am [experimenter's position] and I will be running this experiment today as part of [my/a] student project for the University of York and BBC R&D.

[Give them a consent form]

This is a consent form which provides a brief outline of what will be involved in the experiment. If you have questions about any of it please feel free to ask or if you feel unhappy with participating in this study you are free to withdraw now or at any time throughout the experiment. If you would like to have a read of the form?

The experiment today will be over three sessions, of about 20 minutes each. Between these sessions we will have a 20 minute break in which you can go back to work. In the first of these sessions I will introduce you to the interface and give you some practice and then in the second and third sessions I will ask you to perform the experiment.

The device in front of you is a MIDI controller which I will ask you to use to navigate through a list of words which you will hear using the headphones provided. To start I will ask you to put on the headphones and using this dial [gesturing to appropriate control] set the playback level to one which you can hear the words clearly and find comfortable. Feel free to adjust to adjust the headphones so that they fit you.

[Offer help if they struggle with the headphones. When they are ready, press the trigger button on the interface as they adjust the level]

[Give a copy of the instructions]

This is a set of instructions which explains how to use the interface and the tasks which I am going to ask you to perform today. Please read them and if you have any questions, feel free to ask.

[Allow them to read instructions and answer questions that they ask]

[show them TLX]

This is the rating system interface. You answer each of the questions by marking on the scale at the bottom. You do this by clicking with the mouse and drawing across the line at the relevant point along the scale. If your line crosses the scale multiple times the first point is chosen. If you want to re-draw the line you can by just clicking and dragging again.

[Show them]

To acquaint you with the interface and the rating system I will give you a few practice tasks. In this part of the experiment if you have any questions on how to use the interface feel free to ask them. To get you used to the process the word you selected will be displayed in the bottom box. So if you would like to press the top central pad when you wish to begin the tasks.

[allow the user to complete the task answer any questions they raise]

[When please stop is displayed on the screen direct the user to the TLX and help clarify and confusion with the use of the interface]

[when TLX complete move on to the next trial and repeat.]

You have completed a number of tasks and had some experience with the rating system so we will have a short break now before we move on to the experiment trials in 20 minutes. Other people in the office are likely to also be participating in the experiment, so I would ask you not to discuss the experiment with them, as it could influence the results.

[Allow participant to return to work for 20 minutes]

In this session I am going to ask you to perform tasks as part of the experiment similar to those that you completed in the previous session. Unlike in the previous tasks information about the word you selected will not be displayed. In these tasks we ask that you try to locate and select the target word as quickly as you can whilst trying to be as accurate as possible. Do you have any questions before we continue with the experiment?

[open experimental patch]

[Perform experiment at the ends of each block ask the participant to perform the NASA TLX evaluation as with the training]

[4 blocks of experiment]

Ok thank you that is all of the tasks in this session. We will have another break for 20 minutes before the final session of the testing.

[Allow participant to return to work for 20 minutes]

I am going to ask you again to perform a number of tasks, similar to those you completed within the last session. After this I will have a few questions for you to answer and then the experiment will be complete. If you are ready then please put on your headphones and we will begin.

**After completion of the trials:**

*All participants*

E: Ok, Thank you that is the end of the tasks. I just have a few questions and then the experiment is finished. How did you find completing the tasks today?

*Participants: 0, 2, 4, 6, 8, 10, 12, 14*

If you had to navigate spoken menus on a regular basis, would you prefer to use a display in which the words were overlapping or were presented one at a time?

*Participants: 1, 3, 5, 7, 9, 11, 13, 15*

If you had to navigate spoken menus on a regular basis, would you prefer to use a display in which the words were presented one at a time or overlapping?

*All participants*

When navigating the lists did you develop any listening strategy for finding the target?

*All participants*

When you were completing the tasks, do you feel that you were aware of the identity of the other words in the list?

*Participants: 0, 1, 4, 5, 8, 9, 12, 13*

During the two experimental sessions the words were different lengths. Do you feel that the words being longer made completing the tasks easier, made no difference, or made it more difficult?

*Participants: 2, 3, 6, 7, 10, 11, 14, 15*

During the two experimental sessions the words were different lengths. Do you feel that the words being longer made completing the tasks more difficult, made no difference or made it easier?

Ok, thank you that is the end of the experiment. The purpose of this experiment has been to evaluate different designs of auditory display for the presentation of lists. Between the conditions the amount of delay between the start of the words within each group of three

was varied. We are interested in how presenting overlapping words affects the speed of navigation and the workload for the user and whether this would be an effective approach for the non-visual display of television menus for applications such as programme guides and option lists for users who are unable to see the screen either due to health or situational impairments.

I ask you not to discuss the experiment with other colleagues until we have completed the experimental trials on [insert date/time of end of trials] in case it influences people who will be participating in the experiment later. I have a document here which provides further information on the study and contact information for myself and my supervisors if following this experiment you have any further questions, comments, complaints, or wish to remove your data from the experiment.

[Give them debriefing form]

Thank you very much for your time and your feedback.



## B.3 Consent form

### **Participant consent form: Spoken Auditory Menus**

Investigators: Alistair Hinde, Tony Tew, Dr. Mike Evans, Prof. David Howard

#### **Outline of study:**

The aim of this experiment is to assess how variations in the presentation of spoken lists affects the user experience of the system as well as the accuracy and speed with which users can navigate through the list to reach desired content. We need people to participate in this experiment in order to assess how variations in the design features affects performance, as we cannot evaluate the system ourselves. I will ask you to navigate as quickly as possible through the spoken list of words and select a particular word, which will be displayed to you via a screen, using a set of pads on a MIDI controller. The spoken list will be presented to you using a pair of headphones. As with all presentation of audio, there is a possibility of hearing damage. To minimise this risk the headphones you will use have been limited and you will be asked to set your own playback level (volume) to one which you find comfortable. You will have the ability to adjust the playback level throughout the experiment.

I will be observing as you complete the tasks and your interactions with the interface will be recorded. At points during the experiment I will ask you to rate a number of factors concerning your experience of the interface and at the end of the study I will ask you a few questions regarding your opinions on the interfaces. I will also be making an audio recording of the session so that I have a record of your comments. The data I record will be stored securely in an anonymised form and deleted at the end of the research project.

The purpose of this experiment is to evaluate the systems and not you.

#### **Declaration of consent:**

##### **By agreeing to participate in this experiment I accept that:**

- I have been informed of the procedures, risks and aims involved in the experiment
- This experiment is being performed as part of a research collaboration between the University of York and BBC R&D. My data will be used only for research purposes.
- The data I provide may be published internally or externally and be used as part of presentations related to the research. Any publication of the data will be in an anonymised form with all identifying information removed.
- My participation is entirely voluntary and no remuneration will be provided for my time.
- I am free to withdraw from the experiment at any time. In which case, unless otherwise agreed, the data I provided as part of the study will be destroyed

Please feel free to ask any questions. Then, if you are willing to participate, please sign and date this consent form and proceed with the experiment.

Name in capitals: \_\_\_\_\_ Signature: \_\_\_\_\_

Date: \_\_\_\_\_

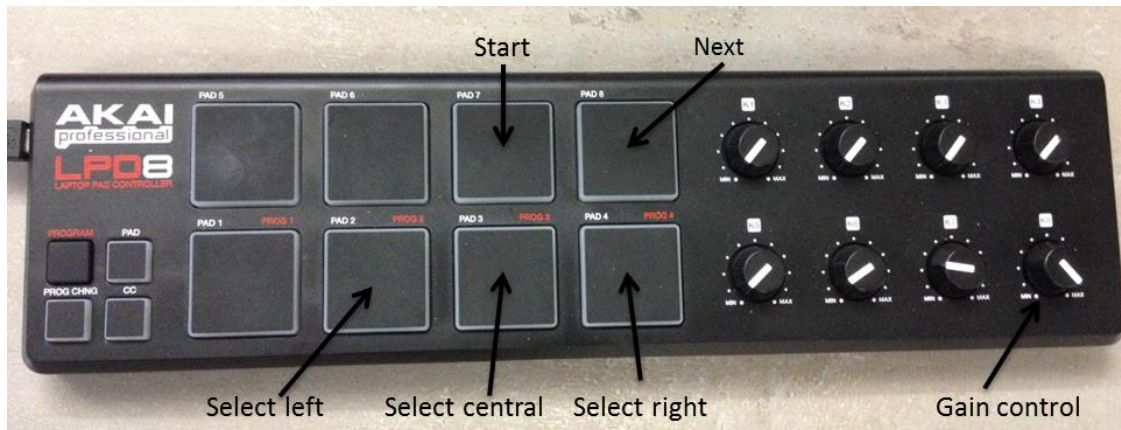
## B.4 Instructions

### Task instructions

For the experiment I will ask you to find and select a specific word in a list. The list will be displayed to you as speech over headphones. The list will be split into groups of three words which will be presented from different spatial location and at different pitches. One word will appear from the left, one from the centre and one from the right. The pitch of the words will increase from left to right.

#### The controls

You can navigate the list using the interface in front of you. Today you will only need to use 5 of the 8 pads to complete the tasks. The pads you will need are labelled: start, next, select left, select centre and select right and are highlighted in the figure below.



To start each task, and display the first items you press the *start* pad. The *next* pad can be used to move forwards in the list to the next group of 3 items. Pressing the *next* pad will begin playback of these items immediately. You will only be able to move forwards in the list and the words will only be played once.

If you hear the target word you can select it by pressing one of the three pads on the bottom row. The positions on the interface relate to the locations from which the targets will be presented (i.e. pressing the *select left* pad will select the word on the left, *select centre* will select the word in the middle and the *select right* will select the right-most word). Alternatively, if you do not hear the target word, pressing the *next* pad will move forwards to the next three items or end the task if it is the end of the list.

#### An outline of the tasks

At the beginning of each task a target word will be shown on-screen and then when you feel ready to start the task you press the central button on the top row, which will play the first sound in the list and allow you to begin navigating the list. The target word will remain on the screen for the duration of the task.

Each list will contain three groups of three words. If you have not heard the target, continuing to navigate forward from the final group of three will end the task. Some lists in the experiment will not include the target word, when this is the case the correct response is to press the 'next' pad when on the final triplet, as if navigating on to the next item.

The pads on this device are not designed to be used in this way and therefore it is best to give them a strong tap rather than pressing them as you would a button. You will see a light behind the button flash when your tap has been registered.

You will notice on the screen with the target word there will be a box with a cross in it. When a task is in progress this cross will disappear and reappear when a selection has been made.

Throughout the experiment you will use several different variations of display. I will ask you to perform a number of tasks with each and then to rate several factors regarding your experience using the interface. The ratings you provide will be combined to create a workload score. Please consider your ratings carefully and try to be consistent with how you marked other factors. Don't think there is any correct pattern: it is your opinions which we are interested in.

If you have any questions, please feel free to ask now.

## B.5 Debriefing form

### Concurrency in Auditory menus

Thank you for your participation in this experiment. Through the experiment you used auditory displays which presented words of different lengths with varying amounts of asynchrony. We are interested in whether there is a particular amount of asynchrony that will allow users accessing TV lists and menus to more quickly navigate accurately and whether this is dependent on the length of the word. Along with the speed possible with these interfaces we are also interested in the workload, which users' experience with these displays and how the amount of asynchrony affects this. The ratings which we asked you to perform are combined for this so that we can assess the relative difficulty of different designs of display.

If you wish to have further information on this project, would like to be kept informed on the results of this study, or wish to remove your data from the study you can contact me on: [Alistair.hinde.ext@bbc.co.uk](mailto:Alistair.hinde.ext@bbc.co.uk)

If you have any concerns or complaints regarding this experiment you can contact my supervisors' either Mike Evans at the BBC ([Michael.evans@bbc.co.uk](mailto:Michael.evans@bbc.co.uk)) or Tony Tew at the University of York ([tony.tew@york.ac.uk](mailto:tony.tew@york.ac.uk)).

Thank you again,

Alistair Hinde

Research student,  
Department of Electronics,  
University of York,  
Based in: BBC R&D

## B.6 Data access

Data from the pilot and main experiments described in Chapter 5 cannot be made available, as sufficient consent was not gathered from participants.



## Appendix C

# Upmixing formula

$$c_m = 0.707106781186548$$

$$s_m = 0.707106781186548$$

$$b = 0.25$$

$$c_1 = 0.25$$

$$c_2 = -0.2$$

$$a_1 = 1 - bc_m - c_1s_m$$

$$a_2 = -bc_m - c_2s_m$$

$$R_f = a_1R_{input} + a_2L_{input}$$

$$L_f = a_1L_{input} + a_2R_{input}$$

$$LFE = 0$$

$$C = b(L_{input} + R_{input})$$

$$R_r = c_1R_{input} + c_2L_{input}$$

$$L_r = c_1L_{input} + c_2R_{input}$$

Where  $L_{input}$  and  $R_{input}$  are the left and right input channels and  $R_f$ ,  $L_f$ ,  $C$ ,  $LFE$ ,  $R_r$  and  $L_r$  are the six channels of the 5.1. Equations are derived from the software (Marston, 2016).





## Appendix D

# Secondary Programme Content Experiment Documents

### D.1 Ethical approval

Ethical approval was granted by the Physical Sciences Ethics Committee of the University of York for the pilot and main experiments on the 20/10/2015 with the reference number: Hinde150928. Following the pilot study, approval was obtained from the ethics committee to change where the data was stored to accommodate the larger file sizes and to modify the exclusion criteria so that only participants that considered themselves to be native English speakers were included in the study.

## D.2 Ethics application

PSEC Application Form V3

THE UNIVERSITY *of York*

### Application Form for Physical Sciences Ethics Committee Approval

***Advice for applicants on completing the form***

*Please ensure that the information provided is:*

- *Accurate and concise*
- *Clear and simple and easily understood by a lay person*
- *Free of jargon, technical terms and abbreviations*

*Further advice and information can be obtained from your departmental representative on the PSEC and at: <http://www.york.ac.uk/admin/aso/ethics/cttee.htm>*

***Please return completed form to your departmental representative:***

Dr Helena Daffern, Department of Electronics

***Title of project:*** *Non-visual Display for Connected Television*

#### **SECTION 1 DETAILS OF APPLICANTS**

**Details of principal investigator (name, appointment and qualifications)**

Alistair Hinde; PhD Student with the Audio Lab Research Group, but based in BBC R&D User Experience Group; MSc, BSc(Hons)

**Names, appointments and qualifications of additional investigators** (*student applicants should include their project supervisor(s) here*)

Anthony Tew; Senior lecturer; BSc, CEng, MIET, MIEEE, AMAES, MIPEM

Prof. David Howard; Head of Department; BSc (Eng), PhD, CEng, FIET, FIOA, MAES

Mike Evans; Research lead in User Experience and Accessibility, BBC R&D; MEng, CEng, DPhil

**Location(s) of project**

BBC R&D (MediaCityUK, Salford) & University of York

**SECTION 2 FUNDERS****What is the funding source(s) for the project?**

ESPRC iCase studentship with University of York and BBC R&D

**Please answer the following:**

- (i) Does the express and direct aim of the research or other activity raise ethical issues?  
 YES  NO  X
- (ii) Is there any obvious or inevitable adaptation of research findings to ethically questionable aims?  
 YES  NO  X
- (iii) Is the work being funded by organisations tainted by ethically questionable activities?  
 YES  NO  X
- (iv) Are there any restrictions on academic freedoms – notably, to adapt and withdraw from ongoing research, and to publish findings?  
 YES  NO  X

If you answered **Yes** to any of the above, please give details below:

**SECTION 3 DETAILS OF PROJECT OR OTHER ACTIVITY****Aims (100 words max)**

This experiment is looking at scenarios in which a user is watching a television program (main program content) and concurrently accessing related information (secondary program content). The experiment intends to compare several auditory presentations of secondary program content with a visual presentation of the same material and explore the effect of manipulations to the main-program audio. We are interested in how these treatments affect factors of the user experience and user behaviour when exposed to these displays.

**Background (250 words max)**

Companion content or second-screen experiences for television have attracted research interest over the last few years. These experiences, however, have been almost exclusively visual. Auditory display has the potential to reduce the load on a user's visual attention while still allowing access to the additional information. This may also be useful for users with visual impairments or who are blind and therefore unable to view the second screen device, or users with low literacy who may struggle to read textual companion content. For these reasons we are interested in comparing spoken representations of secondary content to visual content.

Audio from the main program may impact the users' abilities to understand and pay attention to the secondary content. We are therefore interested in exploring the effects of removing elements from the main program soundtrack during the delivery of the secondary content and the impact of this on user experience.

**Brief outline of project/activity (250 words max)**

The experiment will consist of each participant watching short clips taken from two television programs. The experiment will be a mixed-design, with the presentation method used for the secondary content as a between-subjects factor and the program soundtrack manipulations as a within-participants factor.

There will be four levels of the secondary content presentation: two of which will be audio presented from loudspeakers positioned at different locations around the listener. The other two conditions will present the secondary content from a smartphone, held by the participant, either as audio or on-screen text. There will be four levels of main program soundtrack manipulation which will consist of different combinations of items being muted during the presentation of the secondary content.

Following each clip, participants will be asked to provide a series of ratings to assess the amount of disruption between elements from the main and secondary content, their perceived workload and how much they would like to use each of the display variants. For each clip there will also be an opportunity to provide additional comments. Participants will also be asked to complete a handedness questionnaire (Edinburgh handedness inventory) for demographic information including their age and sex.

Prior to the experiment proper, participants will watch a sequence of clips with secondary content presented, one with each of the presentation methods with unmodified main program audio. During this time they will be asked to consider the rating scales used in the experiment proper to contextualise the ratings they give.

**Study design** (*if relevant – e.g. randomised control trial; laboratory-based*)

The experiment is a mixed design with one between and one mixed-subjects factor each with four levels.

We propose a pilot study to precede the main experiment which will be used to check the factors of the experimental design (i.e. duration, question phrasings, any points of confusion). Pilot participants may be asked additional questions regarding the experimental procedure at the end of the experiment to identify any elements that could be improved. It is also hoped that it will provide an insight into the response patterns we will expect in the main experiment.

PSEC Application Form V3

**If the study involves participants, how many will be recruited?**

For the pilot study (8) participants will be recruited (2 per group).

For the main experiment (32) participants will be recruited (8 per group).

**If applicable, what is the statistical power of the study, i.e. what is the justification for the number of participants needed?**

For the pilot experiment 2 participants per group will provide some indication of the expected trends and the amount by which participants' responses will vary. This will also provide thorough examination of the experimental procedure.

For the main experiment, the minimum number of participants to balance ordering of the within-participant condition is 16. However, it seems likely that 4 participants per group will not provide adequate power. 8 participants per group is the next number that allows the ordering to be accounted for.

**SECTION 4 RECRUITMENT OF PARTICIPANTS****How will the participants be recruited?**

For the pilot participants will be recruited from the BBC.

Participants will be recruited from the students and staff at the University of York

**What are the inclusion/exclusion criteria?**

As part of recruitment participants will be asked to volunteer only if they are right-handed, believe that they have normal-hearing, and normal or corrected to normal vision. Participants will also be asked to confirm these statements on the consent form. If a participant does not confirm all of these statements they will be removed from the experiment, and an additional participant will be recruited.

**Will participants be paid reimbursement of expenses?**

YES NO 

**Will participants be paid?**

YES NO 

*If yes, please obtain signed agreement*

\*Participants of the main experiment will be paid but those in the pilot will not be. See agreement in consent forms and additional payment receipt attached.

**Will any of the participants be students?**

YES NO 19<sup>th</sup> March 2014

**SECTION 5 DATA STORAGE AND TRANSMISSION**

**If the research will involve storing personal data, including sensitive data, on any of the following please indicate so and provide further details (answers only required if *personal* data is to be stored).**

<b>Manual files</b>	Consent forms and payment receipts will be stored in a locked draw at the BBC.
<b>University computers</b>	
<b>Home or other personal computers</b>	
<b>Laptop computers, tablets</b>	All participant data will be stored on an encrypted BBC laptop.  The recorded video data will be transferred to this laptop and deleted from the camera's memory.  An encrypted external drive may also be used to store data as backup and allow transfer data between machines for analysis.
<b>Website</b>	The data-collected from the main experiment will be published via the University of York in an anonymised form (to comply with EPSRC expectations). Videos and paperwork, however, will not be made available due to the identifiability of participants.

**Please explain the measures in place to ensure data confidentiality, including whether encryption or other methods of anonymisation will be used.**

All of the digital data will be stored on encrypted devices during the project. Video data cannot be anonymised. Participant response data and videos will be referred to by the same participant numbers for analysis. Any publication of data, however, will be anonymised. Consent and payment receipts cannot be anonymised and therefore will be stored securely in a locked draw.

**Please detail who will have access to the data generated by the study.**

The persons noted as investigators. (Data will be published and may be used by other researchers)

PSEC Application Form V3

**Please detail who will have control of and act as custodian for, data generated by the study.**

Alistair Hinde

**Please explain where, and by whom, data will be analysed.**

Data will be analysed at BBC R&D in MediaCityUK by Alistair Hinde, with possible input from the other investigators and members of the User Experience Research Group at the BBC. As the data will be published, further analysis may be performed by other researchers.

**Please give details of data storage arrangements, including where data will be stored, how long for, and in what form.**

During the project, the digital data will be stored on Alistair's encrypted BBC laptop, on an encrypted external drive.

Anonymised data from the main experiment (excluding the videos + paper forms) will be published online using the university system to meet EPSRC expectations. To meet these expectations this data may be stored indefinitely and will be available for other researchers to use. Data from the pilots and the videos from the experimental sessions, however, will be deleted at the end of the research project (expected April 2016).

Consent and payment forms will be kept in a locked draw at the BBC for the duration of the project, after which they will be destroyed.

19<sup>th</sup> March 2014



**SECTION 6 CONSENT****Is written consent to be obtained?**YES NO *If yes, please attach a copy of the information for participants*

Two consent forms accompany this submission. One version is for the participants of the pilot and one is for the main experiment to account for differences in the data handling and scope of the experiments.

*If no, please justify***Will any of the participants be from one of the following vulnerable groups?**

Children under 18

YES	<input type="checkbox"/>	NO	<input checked="" type="checkbox"/>
-----	--------------------------	----	-------------------------------------

People with learning difficulties

YES	<input type="checkbox"/>	NO	<input checked="" type="checkbox"/>
-----	--------------------------	----	-------------------------------------

People who are unconscious or severely ill

YES	<input type="checkbox"/>	NO	<input checked="" type="checkbox"/>
-----	--------------------------	----	-------------------------------------

People with mental illness

YES	<input type="checkbox"/>	NO	<input checked="" type="checkbox"/>
-----	--------------------------	----	-------------------------------------

NHS patients

YES	<input type="checkbox"/>	NO	<input checked="" type="checkbox"/>
-----	--------------------------	----	-------------------------------------

Other vulnerable groups (if 'yes', please give details)

YES	<input type="checkbox"/>	NO	<input checked="" type="checkbox"/>
-----	--------------------------	----	-------------------------------------

**If so, what special arrangements have been made for getting consent?**

n/a

**SECTION 7 DETAILS OF INTERVENTIONS****Indicate whether the study involves procedures which:**

Involve taking bodily samples

YES	<input type="checkbox"/>	NO	<input checked="" type="checkbox"/>
-----	--------------------------	----	-------------------------------------

Are physically invasive

YES	<input type="checkbox"/>	NO	<input checked="" type="checkbox"/>
-----	--------------------------	----	-------------------------------------

Are designed to be challenging/disturbing (physically or psychologically)

YES	<input type="checkbox"/>	NO	<input checked="" type="checkbox"/>
-----	--------------------------	----	-------------------------------------

**If so, please list those procedures to which participants will be exposed:****List any potential hazards:**

- (i) Hearing damage: as with any sound exposure. Where possible the system will be calibrated to ensure presentation level is within safe levels for the durations required for the experiment. As the level at the ear is dependent on the distance from the source and the participant will have control over this in the 'held' condition, it is not possible control the exposure in this condition. In the consent, participants will be notified of this potential risk and advised to position the phone's speaker directly to their ear. This is self-regulated, however, and therefore it is assumed that participants will not choose to listen in a way which causes discomfort or damage. Participants will

be reminded of the risks of overexposure to noise and their ability to withdraw should they experience any discomfort will be outlined in the consent form.

- (ii) There is also a risk of electrocution from the use of electronics. The possibility of any harm coming to the participant will be minimised by ensuring all devices running from mains power have been PAT tested.
- (iii) There is a risk of tripping due to the requirement of cables. However, the researcher will minimise this risk by routing cables so that they are out of the way where possible and taping down any loose cables.

**List any discomfort or distress:**

Possible discomfort with audio presentations see item (i) in potential hazards.

**What steps will be taken to safeguard**

- (i) the confidentiality of information

Data will be stored securely as detailed in section 5. Only researchers specified will be involved in the analysis non-anonymised video data for analysis.

- (ii) the specimens themselves?

No specimens will be taken

**What particular ethical problems or considerations are raised by the proposed study?**

Consideration of presentation levels (item (i) in potential hazards) and data handling (precautions outlined in section 5)

**What do you anticipate will be the output from the study? Tick those that apply:**

- Peer-reviewed publications
- Non-peer-reviewed publications
- Reports for sponsor
- Confidential reports
- Presentation at meetings
- Press releases

**Is there a secrecy clause to the research?**

YES

NO

*If yes, please give details below*

**SECTION 8 SIGNATURES**

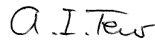
The information in this form is accurate to best of my knowledge and belief and I take full responsibility for it.

I agree to advise of any adverse or unexpected events that may occur during this project, to seek approval for any significant protocol amendments and to provide interim and final reports. I also agree to advise the Ethics Committee if the study is withdrawn or not completed.

Signature of Investigator(s):



(Alistair Hinde, Date: 04/09/2015)



(Anthony Tew, Date: 10/09/2015)

***Responsibilities of the Principal Researcher following approval***

- If changes to procedures are proposed, please notify the Ethics Committee
- Report promptly any adverse events involving risk to participants

## D.3 Consent For the main experiment

### **Participant Consent Form: Secondary Content Presentation**

*Investigators: Alistair Hinde<sup>1</sup>, Tony Tew<sup>1</sup>, David Howard<sup>1</sup> and Michael Evans<sup>2</sup>  
(<sup>1</sup> University of York, <sup>2</sup>BBC Research and Development)*

The aim of this experiment is to evaluate different methods of presenting additional 'secondary' content to TV programs. We need people to participate in this experiment in order to assess how the designs affect the experience for users. Today, I will ask you to answer a few demographic questions and fill out a questionnaire which will be used to assess your handedness (how right or left handed you are). You will be asked to watch clips from two television programs with added secondary content. At points in the experiment you will be asked to answer questions related to your experience of the clips with the different presentation methods. In addition to this we will be videoing the session as we are also interested in how users behave during the experiences.

Before continuing to the experiment, we need to check that you consider yourself right handed, believe your hearing to be normal and that your eyesight is normal or corrected to normal. We request, therefore, that you end the experiment at this point if you are unable to confirm these points. If you normally wear glasses when watching television, we ask that you wear them during the program clips.

As with all presentation of audio, there is a possibility of hearing damage. To mitigate these risks, where possible we have set the levels so as to be safe for the durations of this experiment. There will be times when you are asked to hold a smartphone which will make sounds. In these conditions it is not possible for us to control the level of the sound at your ear. We ask, therefore, that you hold the device at a distance where you can comfortably hear it and avoid positioning the speaker directly to your ear. If at any point during the experiment you find the playback level to be uncomfortable, please inform the assistant and the experiment will be terminated.

The data you supply will be used for research purposes and analyses of this data, and quotes from any comments may be published. The video recordings will be stored and analysed without anonymisation (i.e. blurring). However, this data will be stored securely until the end of this research project (expected 2016) at which point it will be destroyed. The forms you are asked to sign today will also be kept until the end of the project when they will be destroyed. The rest of the data you supply during the experiment may be stored indefinitely and shared with other researchers, in which case they may perform their own analyses on the data. Any data that is shared or published will be anonymised.

The purpose of this experiment is to evaluate the systems and not you. You are free to end the experiment at any time if you no longer wish to continue. In the event that you wish to end the experiment, you are not required to provide any reason.

**Declaration of consent:**

By agreeing to participate in this experiment I accept that:

- I have been informed of the procedures and risks involved in the experiment.
- This experiment is being performed as part of research collaboration between the University of York and BBC R&D. My data will be used only for research purposes.
- I have been informed of how my data will be treated and I consent to its usage as has been outlined in the document above.
- I will receive a payment of £5 for my participation in this study, but no further remuneration will be provided.
- If at any point during the experiment I no longer wish to participate in this study, I can withdraw by informing the researcher running the session. In this eventuality I am not required to provide any reason.

I also confirm that I consider myself:

- To be right handed
- To have normal hearing
- To have normal or corrected to normal vision
- To be a native English speaker

Please feel free to ask any questions. Then, if you are willing to participate in the experiment, please indicate that you agree to the terms outlined in this form by providing your signature below. Otherwise please inform the assistant that you wish to end the experiment.

Printed: \_\_\_\_\_

Signed: \_\_\_\_\_

Date: \_\_\_\_\_

## D.4 Instructions before training

In the experiment today we will present you with a sequence of clips from television programs. Accompanying each of these clips will be some “secondary content”, which will add extra information relating to the show. The experiment is split into two parts. In the first part of the experiment we will show you clips from the program “Upstairs Downstairs” with some different ways of displaying the secondary content. This part of the experiment is just to introduce you to how this content could be displayed and give you some practice thinking about some of the factors that we will ask you about in the second part of the experiment.

For each of the clips in the first part of the experiment the secondary content will be displayed in a different way. In some of the clips the secondary content will be presented as audio from one of the loudspeakers or from the phone that you will be given, or alternatively as text on the screen of the phone. The only speakers that will be used in today’s experiment are the ones on the floor stands that are positioned around you and speaker in the phone. Please note, therefore, that the speakers attached to the ceiling will not be used in this experiment.

This experiment is intended to simulate an interactive system in which a short audio notification informs you that some new secondary content is available. In actual use, if you wished to access the content you would then perform a simple action to display it. In this experiment, however, the system is not interactive and secondary content will be automatically displayed following the notification sound. The notification and the secondary content will be presented from the same location with a short delay to separate them.

When you hear the notification, imagine that you perform a sideways swipe on the screen of the secondary device (i.e. the mobile phone) to trigger the secondary content. If you wish, you may find it useful to physically enact this gesture on hearing the notification during the experiment.

Before the start of each clip you will be informed how the secondary content will be presented. This information will appear on the television screen in front of you, in a similar way to what is being displayed now. If you are unclear about how you should be expecting the secondary content to be presented please ask the person running the experiment to explain.

In clips where the information will be presented from the phone (visually or sonically) please feel free to handle the phone. In the textual display you may have to scroll downwards to see all of the text. This can be done by pressing one finger on the screen and dragging towards the top of the screen. When handling the phone please try not to press any of the buttons

or exit the app.

In front of you there is a sheet of paper with a series of questions and rating scales on it. During this session we just want you to think about these questions and at the end of each clip just consider how you would rate each of the questions. You will not be asked to respond to these questions in this practice session they are simply intended to get you used to thinking about these factors. Please take a moment now to look through these questions. If you have any queries please ask now.

When you are happy to continue with the experiment please tell the experimenter and they will start the first clip. <sup>[1]</sup>

## D.5 Instructions before experiment trials

As with the instructions before the training, minor updates were made to the instructions following the first two participants.

### D.5.1 Front group

During this part of the experiment you will see clips from the program “Lost Land of the Jaguar”. In the practice session just now, you experienced a series of different secondary content presentations. Now, we will again ask you to watch a sequence of clips but this time the secondary content will always be presented from the centre-front speaker (see diagram on the television screen). As before, each piece of secondary content will be preceded by a short notification from the location where the secondary content will come from.

At the end of each clip we will ask you to respond to the set of questions we introduced to you in the practice session. When you are rating the questions, try to keep in mind the practice presentations. There will also be an opportunity for you to provide any additional information about your experience of the presentation.

### D.5.2 Side group

During this part of the experiment you will see clips from the program “Lost Land of the Jaguar”. In the practice session just now, you experienced a series of different secondary

---

<sup>[1]</sup>The first two participants received a slightly different version of the instructions. Changes included information about which of the loudspeakers visible in the room were involved in the experiment. The changes were considered to have an insignificant impact on participants’ responses

content presentations. Now, we will again ask you to watch a sequence of clips but this time the secondary content will always be presented from the speaker to your left (see diagram on the television screen). As before, each piece of secondary content will be preceded by a short notification from the location where the secondary content will come from.

At the end of each clip we will ask you to respond to the set of questions we introduced to you in the practice session. When you are rating the questions, try to keep in mind the practice presentations. There will also be an opportunity for you to provide any additional information about your experience of the presentation.

### **D.5.3 SD-A group**

During this part of the experiment you will see clips from the program “Lost Land of the Jaguar”. In the practice session just now, you experienced a series of different secondary content presentations. Now, we will again ask you to watch a sequence of clips but this time the secondary content will always be presented from the smartphone as sound (see diagram on the television screen). As before, each piece of secondary content will be preceded by a short notification from the location where the secondary content will come from.

At the end of each clip we will ask you to respond to the set of questions we introduced to you in the practice session. When you are rating the questions, try to keep in mind the practice presentations. There will also be an opportunity for you to provide any additional information about your experience of the presentation.

### **D.5.4 SD-V group**

During this part of the experiment you will see clips from the program “Lost Land of the Jaguar”. In the practice session just now, you experienced a series of different secondary content presentations. Now, we will again ask you to watch a sequence of clips but this time the secondary content will always be presented from the smartphone as text (see diagram on the television screen). As before, each piece of secondary content will be preceded by a short notification from the location where the secondary content will come from.

At the end of each clip we will ask you to respond to the set of questions we introduced to you in the practice session. When you are rating the questions, try to keep in mind the practice presentations. There will also be an opportunity for you to provide any additional information about your experience of the presentation.



## D.6 Debriefing

Thank you for completing the experiment today. The purpose of this experiment has been to evaluate the effects on the user experience of presenting additional auditory content from different locations and manipulations to the main program audio. During the second part of today's experiment you watched programs with additional audio content presented from one source. In this experiment different participants will have watched the same clips with the same secondary content with one of the other presentation methods that you experienced in the practice phase of the experiment.

We are particularly interested in the effect of the different conditions on the amount of disruption you experience, the workload associated with each method and how much people like each one. We were videoing the experimental session today because we are also interested in how users behave in the different conditions. With the conditions in which the phone presented the information, either as sound or text, we are interested in how users chose to position the phone throughout the experiment. Also, in all of the conditions we are looking to see where participant looked during the presentations and if they displayed any behaviours that may be associated with exerting effort to focus on either of the content streams.

I ask you not to discuss the details of this experiment with others who may also be participating, as it could influence the way they behave in the experiment.

If you have any questions, comments or complaints please feel free to contact me or my supervisors (contact details below).

Thank you again for your time and feedback.

Alistair Hinde: afh508@york.ac.uk

Tony Tew: tony.tew@york.ac.uk

Mike Evans: Michael.evans@bbc.co.uk<sup>[2]</sup>

## D.7 Response application screenshots

To give an impression of the manner in which participants were prompted to provide responses for the experiment outlined in Section 7.4, screenshots of the response application are provided.

---

<sup>[2]</sup>The first two participants received a different debriefing statement. As this was provided after data collection, it was not considered to have influenced their data

**DEMOGRAPHICS**

Please enter your age:

Please indicate your sex:  Female  Male

Quit Next

**Figure D.1:**  
*The page on which participants entered their demographic data.*

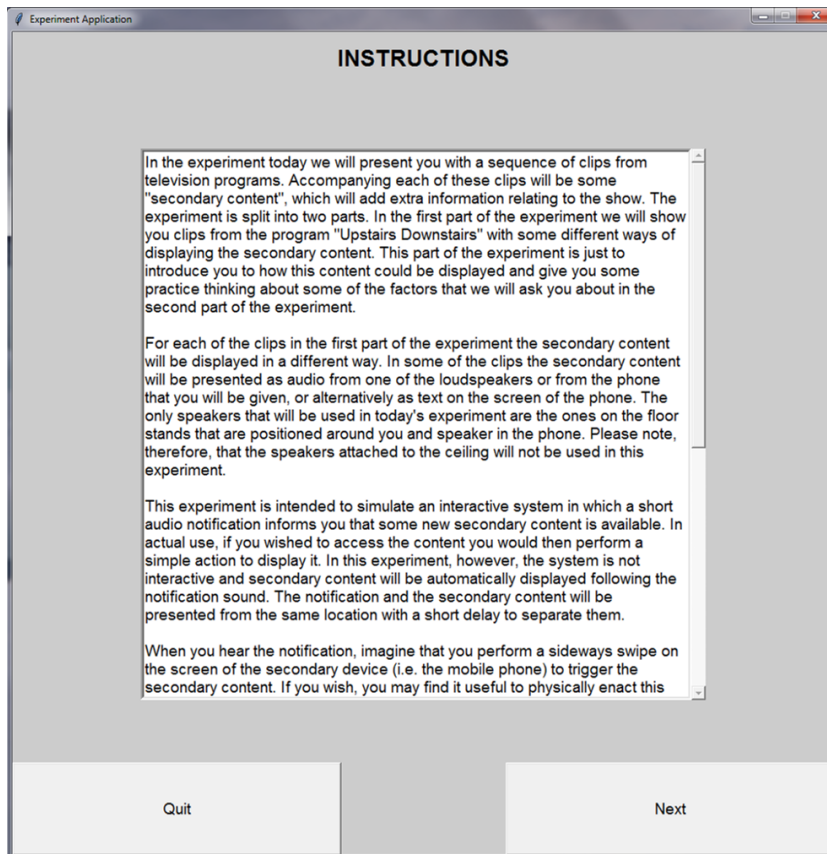
**HANDEDNESS**

Please indicate your preferences in the use of hands in the following activities by clicking in the appropriate column. Where the preference is so strong that you would never try to use the other hand unless absolutely forced to, mark the 'Only Left' or 'Only Right' options. Some of the activities require both hands. In these cases the part of the task, or object, for which hand preference is wanted is indicated in brackets. Please try to answer all the questions, and only mark 'Don't know' if you have no experience at all of the object or task.

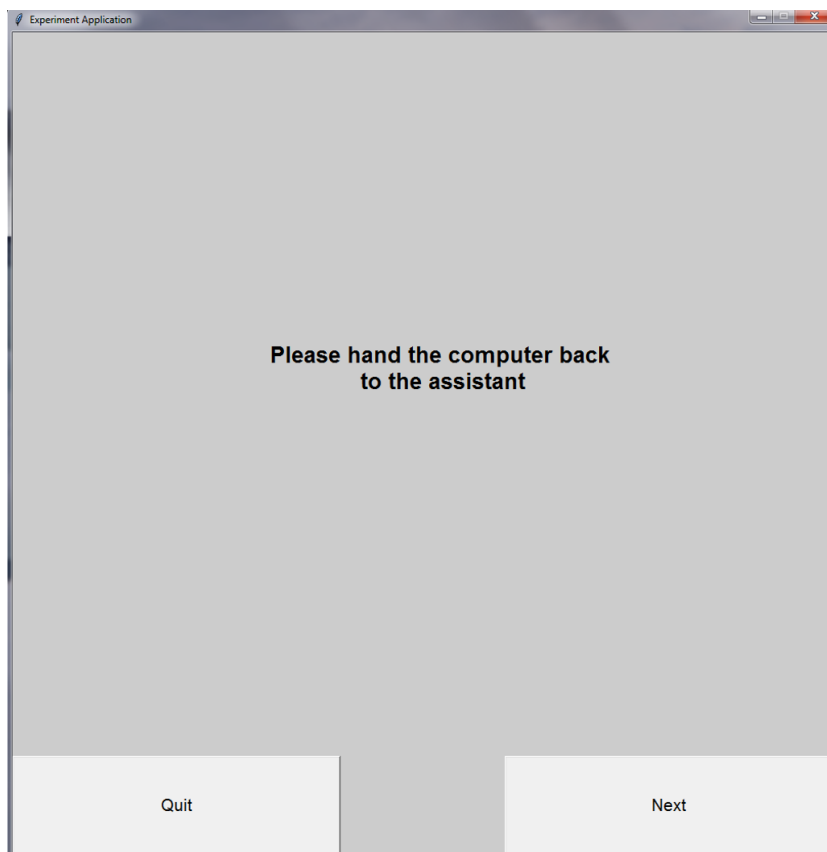
	Only Left	Left	No preference	Right	Only Right	Don't know
Writing:	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Drawing:	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Throwing:	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Scissors:	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Toothbrush:	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Knife (without fork):	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Spoon:	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Broom (upper hand):	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Striking match (match):	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Opening box (lid):	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Quit Next

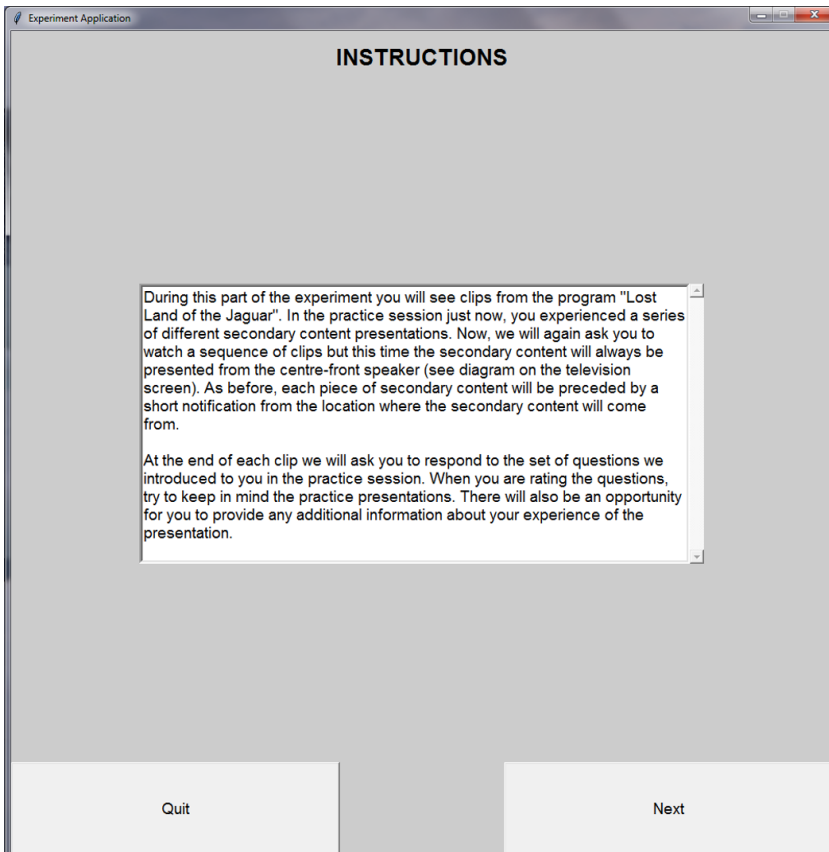
**Figure D.2:**  
*The page on which participants completed the EHI. Wording was chosen to be as close as possible to that used by Oldfield (1971).*

**Figure D.3:**

*The instruction page shown prior to the familiarisation clips. (See full text in Section D.4).*

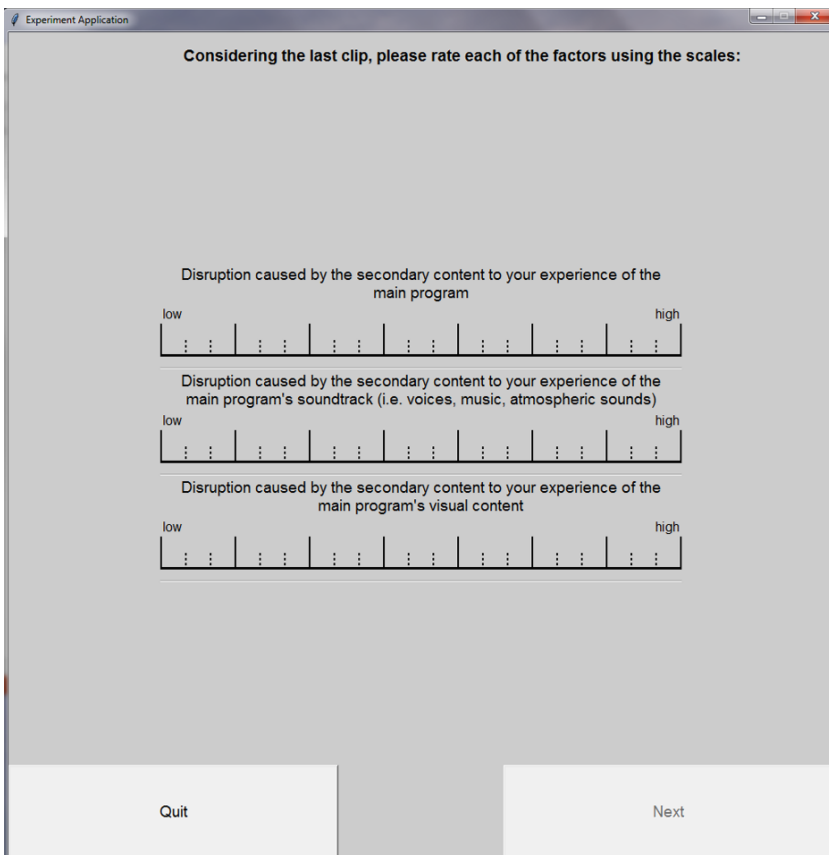
**Figure D.4:**

*This page was used on several occasions when the user had finished a set of tasks with the response application. It followed the first set of instructions (Figure D.3), the second set of instructions (Figure D.5), and after the comments page in all but the final experimental trial (Figure D.10).*



**Figure D.5:**

*The page that presented the user with instructions for the experimental session. It was presented when the user had finished watching the familiarisation clips. See Section D.5 for the instructions used for all of the groups.*



**Figure D.6:**

*The first page of disruption ratings. This page and subsequent question pages were presented after each of the experimental clips. The analysis of which is described in Section 7.4.5.*

Experiment Application

Considering the last clip, please rate each of the factors using the scales:

Disruption caused by the main program to your experience of the secondary content

low high

Disruption caused by the main program's soundtrack (i.e. voices, music, atmospheric sounds) to your experience of the secondary content

low high

Disruption caused by the main program's visual content to your experience of the secondary content

low high

Quit Next

**Figure D.7:**  
*The second page of disruption ratings. The analysis of which is described in Section 7.4.5.*

Experiment Application

Considering the last clip, please answer the questions using the scales:

How mentally demanding was the experience?

low high

How physically demanding was the experience?

low high

How hurried or rushed was the pace of the experience?

low high

How hard did you have to work during the experience?

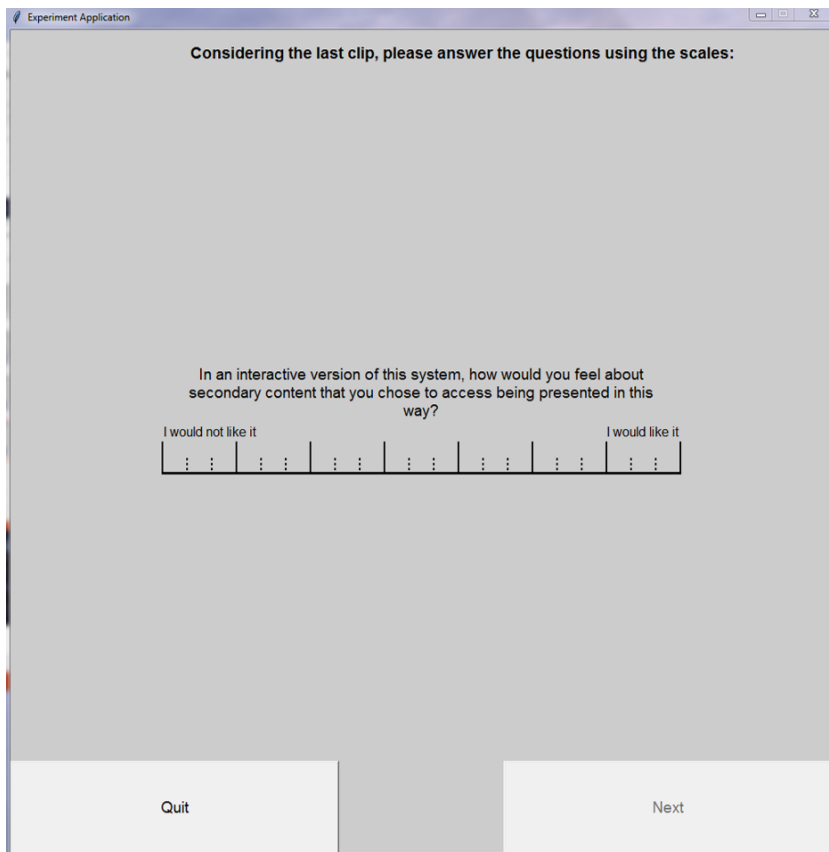
low high

How insecure, discouraged, irritated, stressed, and annoyed were you?

low high

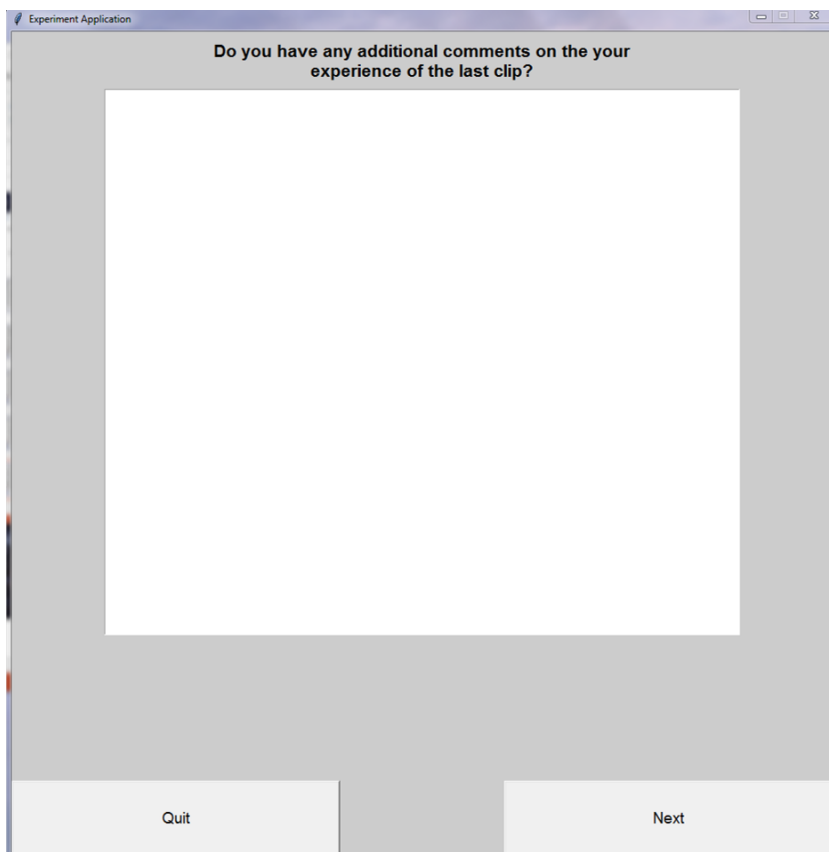
Quit Next

**Figure D.8:**  
*The page on which participants gave the workload ratings. The analysis of which is described in Section 7.4.5.*



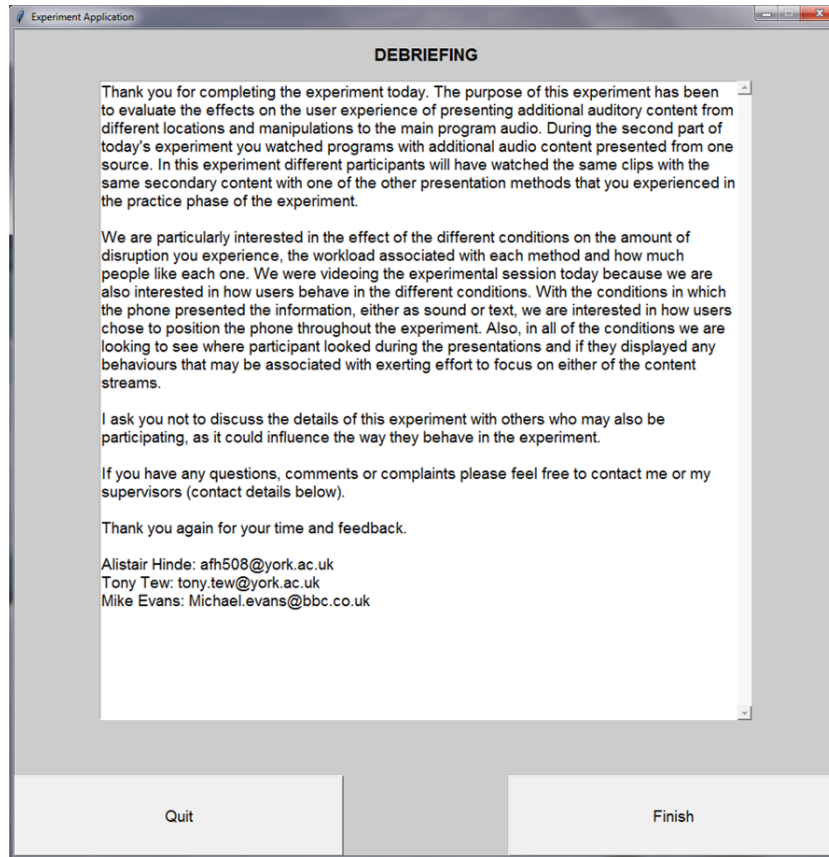
**Figure D.9:**

*The page on which participants gave the preference rating. The analysis of which is described in Section 7.4.5.*



**Figure D.10:**

*The page on which participants gave comments on each of the experimental clips. The analysis of which is described in Section 7.4.6.*



*Figure D.11: The page which displayed the debriefing information to the participants at the end of the experiment. See Section D.6 for a larger type version of the debriefing message and related notes.*

## D.8 Data access

Data from the pilot study outlined in Chapter 7 cannot be made available as sufficient consent was not obtained.

Data from the main experiment (described in Chapter 7) is accessible from DOI: 10.15124/19b43aa8-744a-404b-a8ff-7b0f931bf6d3.





# Glossary of Terms

asynchronous experience	An additional media activity that is experienced at a time other than during the specific programme.
audio description	Spoken descriptions of visual elements added to television programmes or film. Also known as video descriptions.
companion content	Content that has been created or curated as a result of a specific television programme.
evolving	Where the content in an orchestrated experience has not been entirely pre-determined by the orchestrator (e.g., aggregation of current social media activity).
fixed	Where the content in an orchestrated experience is pre-determined by the orchestrator to provide a time-invariant experience.
improvised	A synchronous experience in which the additional media activity has not been created with specific consideration of its use alongside the main programme.
main programme content	All elements of a programme that are intrinsic to conveying the narrative. This includes accessibility features such as audio description and subtitles.
non-companion content	Content that is considered to be related to the television programme at the time of access, but was not created or curated as a result of the show.
orchestrated	An experience that has been facilitated by the content-provider, or a third-party with synchronous use in mind.

programme selection menu	A menu that presents the user with a plurality of television programmes that may be selected to watch and/or record.
secondary programme content	Additional content that is displayed as part of a scheduled or partially-scheduled orchestrated synchronous companion experience.
service selection menu	A menu that presents the user with a choice of television services (e.g., applications).
synchronous experience	An additional media activity that is experienced at the same time as a specific programme.
utility menu	A menu that provides configuration options for the device or application.

# Abbreviations

AD	audio description.
ANOVA	analysis of variance.
ASPC	auditory secondary programme content.
AUDELTEL	AUdio DEscribed TELevision.
BRM	backward recognition masking.
CI	confidence interval.
CRM	coordinate response measure.
CVC	consonant-vowel-consonant.
CVCV	consonant-vowel-consonant-vowel.
dB	decibels.
EHI	Edinburgh handedness inventory.
EPG	electronic programme guide.
ERB	equivalent rectangular band.
fps	frames-per-second.
GEE	generalised estimating equations.
GUI	graphical user interface.
HCI	human-computer interaction.
HRIR	head-related impulse response.
HRTF	head-related transfer function.
Hz	Hertz.
IED	interaural envelope difference.
ILD	interaural level difference.
IOT	internet of things.
IP	internet protocol.
IPA	International Phonetic Alphabet.
IPD	interaural phase difference.

IPTV	internet-protocol TV.
ITD	interaural time difference.
kHz	kilohertz.
LEA	left-ear advantage.
LKFS	loudness, K weighted, relative to nominal full scale.
LUFS	loudness units referenced to full scale.
MIDI	musical instrument digital interface.
MPC	main programme content.
ms	milliseconds.
OBB	object-based broadcasting.
OS	operating system.
OTT	over-the-top.
PC	personal computer.
PDA	personal display assistant.
REA	right-ear advantage.
RF	radio frequency.
rm-ANOVA	repeated measures analysis of variance.
RMS	root mean square.
RNIB	Royal National Institute of Blind People.
SIID	situationally-induced impairments and disabilities.
SNR	signal-to-noise ratio.
SOA	stimuli onset asynchrony.
SPC	secondary programme content.
SRM	spatial release from masking.
STB	set-top box.
TMR	target-to-masker ratio.
TTS	text-to-speech.
UK	United Kingdom.
USB	universal serial bus.
VC	vowel-consonant.
VOD	video on-demand.
VR	virtual reality.
VSPC	visual secondary programme content.

# References

- 2nd Screen Society (n.d.), 'Lexicon for the 2nd Screen Society', [online], Available: <http://www.2ndscreenociety.com/lexicon/#CompanionExp>, [Accessed: 2013-12-27].
- Access Economics Pty Limited (2009), 'Future Sight Loss UK (1): The Economic Impact of Partial Sight and Blindness in the UK Adult Population', Technical Report July, Royal National Institute for Blind People, [online], Available: [http://www.rnib.org.uk/sites/default/files/FSUK\\_Report.pdf](http://www.rnib.org.uk/sites/default/files/FSUK_Report.pdf), [Accessed: 2015-08-09].
- Alais, D. & Burr, D. (2004), 'The ventriloquist effect results from near-optimal bimodal integration', *Current Biology*, 14(3), pp. 257–262.
- Allen, K., Carlile, S. & Alais, D. (2008), 'Contributions of talker characteristics and spatial location to auditory streaming', *The Journal of the Acoustical Society of America*, 123(3), pp. 1562–70.
- Andersen, T. S., Tiippana, K., Laarni, J., Kojo, I. & Sams, M. (2009), 'The role of visual spatial attention in audiovisual speech perception', *Speech Communication*, 51(2), pp. 184–193.
- Android (n.d.), 'Multi-Window Support', [online], Available: <https://developer.android.com/preview/features/multi-window.html>, [Accessed: 2016-08-19].
- Apple Inc. (n.d.a), 'iOS 9 - What's New', [online], Available: <http://www.apple.com/uk/ios/whats-new/>, [Accessed: 2016-08-19].
- Apple Inc. (n.d.b), 'Siri', [online], Available: <http://www.apple.com/uk/ios/siri/>, [Accessed: 2016-01-17].
- Arbogast, T. L. & Kidd, G. (2000), 'Evidence for spatial tuning in informational masking using the probe-signal method', *The Journal of the Acoustical Society of America*, 108(4), pp. 1803–1810.

- Armstrong, G. B., Boiarsky, A. & Mares, M.-L. (1991), 'Background television and reading performance', *Communication Monographs*, 58(3), pp. 235–253.
- Armstrong, G. B. & Chung, L. (2000), 'Background television and reading memory in context: Assessing TV interference and facilitative context effects on encoding versus retrieval processes', *Communication Research*, 27(3), pp. 327–352.
- Armstrong, M., Barrett, J. & Evans, M. (2010), 'Enabling and Enriching Broadcast Services by Combining IP and Broadcast Delivery', Technical report, BBC Research & Development, [online], Available: <http://www.bbc.co.uk/rd/publications/whitepaper185>, [Accessed: 2013-11-18].
- Armstrong, M., Brooks, M., Churnside, T., Evans, M., Melchoir, F. & Shotton, M. (2014), 'Object-based Broadcasting - Curation, Responsiveness and User Experience', Technical report, BBC Research & Development, [online], Available: <http://www.bbc.co.uk/rd/publications/whitepaper285>, [Accessed: 2015-10-19].
- Arons, B. (1997), 'SpeechSkimmer: a system for interactively skimming recorded speech', *ACM Transactions on Computer-Human Interaction*, 4(1), pp. 3–38.
- Assmann, P. F. (1995), 'The role of formant transitions in the perception of concurrent vowels', *The Journal of the Acoustical Society of America*, 97(1), pp. 575–584.
- Assmann, P. F. (1996), 'Modeling the perception of concurrent vowels: Role of formant transitions', *The Journal of the Acoustical Society of America*, 100(2 pt. 1), pp. 1141–52.
- Assmann, P. F. (1999), 'Fundamental frequency and the intelligibility of competing voices', in: '14th International Congress of Phonetic Sciences', pp. 179–182.
- Assmann, P. F. & Summerfield, Q. (1990), 'Modeling the perception of concurrent vowels: vowels with different fundamental frequencies', *The Journal of the Acoustical Society of America*, 88(2), pp. 680–97.
- Asutay, E. & Västfjäll, D. (2015), 'Attentional and emotional prioritization of the sounds occurring outside the visual field', *Emotion*, 15(3), pp. 281–286.
- Avila, C., Furnham, A. & McClelland, A. (2012), 'The influence of distracting familiar vocal music on cognitive performance of introverts and extraverts', *Psychology of Music*, 40(1), pp. 84–93.

- Backhouse, A., Ferguson, L. & Young, J. (2008), 'Lost Land of the Jaguar', [Television Programme], Series: 1, Episode: 1, British Broadcasting Corporation.
- Baddeley, A. (1997), *Human Memory: Theory and Practice*, Psychology Press Ltd., Hove, UK, revised edition.
- Barnicle, K. (2000), 'Usability testing with screen reading technology in a Windows environment', in: 'Proceedings on the 2000 Conference on Universal Usability - CUU '00', pp. 102–109.
- Barrett, J., Hammond, M. & Jolly, S. (2011a), 'The Universal Control Project: An Overview', Technical report, BBC Research & Development, [online], Available: <http://www.bbc.co.uk/rd/publications/whitepaper193>, [Accessed: 2016-09-16].
- Barrett, J., Hammond, M. & Jolly, S. (2011b), 'The Universal Control Project Control API V.0.6.0', Technical report, BBC Research & Development, [online], Available: <http://www.bbc.co.uk/rd/publications/whitepaper194>, [Accessed: 2016-09-16].
- Basapur, S., Harboe, G., Mandalia, H., Novak, A., Vuong, V. & Metcalf, C. (2011), 'Field trial of a dual device user experience for iTV', in: 'Proceedings of the 9th International Interactive Conference on Interactive Television', EuroITV '11, pp. 127–136.
- Basapur, S., Mandalia, H., Chaysinh, S., Lee, Y., Venkitaraman, N. & Metcalf, C. (2012), 'FANFEEDS: Evaluation of a socially generated information feed on second screen as a TV show companion', in: 'Proceedings of the 10th European Conference on Interactive TV and Video - EuroITV '12', pp. 87–96.
- BBC (2013a), 'CEEFAX: World's First Teletext Service, 23 September 1974', 'History of the BBC', [online], Available: <http://www.bbc.co.uk/programmes/p01k3l1q>, [Accessed: 2016-06-02].
- BBC (2013b), 'What is Connected TV?', 'Webwise', [online], Available: <http://www.bbc.co.uk/webwise/0/22728226>, [Accessed: 2016-05-14].
- BBC (2016), 'Planet Earth II', [online], Available: <http://www.bbc.co.uk/programmes/p02544td>, [Accessed: 2017-01-22].
- BBC Research & Development (n.d.a), 'Binaural Broadcasting', 'BBC Research and Development Blog', [online], Available: <http://www.bbc.co.uk/rd/projects/binaural-broadcasting>, [Accessed: 2016-07-20].

- BBC Research & Development (n.d.b), 'Unconventional Screens', [online], Available: <http://www.bbc.co.uk/rd/projects/unconventional-screens>, [Accessed: 2015-10-28].
- BBC Research & Development (n.d.c), 'Visual Perceptive Media', [online], Available: <https://www.youtube.com/watch?v=iXKu-b4Afh4>, [Accessed: 2016-06-06].
- Benford, S., Giannachi, G., Koleva, B. & Rodden, T. (2009), 'From interaction to trajectories', in: 'Proceedings of the 27th International Conference on Human Factors in Computing Systems - CHI 09', pp. 709–718.
- Benjamin, E. (2004), 'Preferred listening levels and acceptance windows for dialog reproduction in the domestic environment', in: 'Audio Engineering Society Convention 117', Article no. 6233, [online], Available: <http://www.aes.org/e-lib/browse.cfm?elib=12890>, [Accessed: 2015-04-14].
- Bentin, S., Kutas, M. & Hillyard, S. A. (1995), 'Semantic processing and memory for attended and unattended words in dichotic listening: Behavioral and electrophysiological evidence', *Journal of Experimental Psychology: Human Perception and Performance*, 21(1), pp. 54–67.
- Bernhaupt, R. & Pirker, M. M. (2014), 'User interface guidelines for the control of interactive television systems via smart phone applications', *Behaviour & Information Technology*, 33(8), pp. 784–799.
- Bertelson, P., Vroomen, J., De Gelder, B. & Driver, J. (2000), 'The ventriloquist effect does not depend on the direction of deliberate visual attention', *Perception & Psychophysics*, 62(2), pp. 321–332.
- Bertelson, P., Vroomenti, J. & de Gelderti, B. (1997), 'Auditory-visual interaction in voice localization and in bimodal speech recognition: The effects of desynchronization', in: 'ESCA Workshop on Audio-Visual Speech Processing (AVSP'97)', pp. 97–100.
- Best, V., Gallun, F. J., Ihlefeld, A. & Shinn-Cunningham, B. G. (2006), 'The influence of spatial separation on divided listening', *The Journal of the Acoustical Society of America*, 120(3), pp. 1506–1516.
- Best, V., Mason, C. R. & Kidd, G. (2011), 'Spatial release from masking in normally hearing and hearing-impaired listeners as a function of the temporal overlap of competing talkers', *The Journal of the Acoustical Society of America*, 129(3), pp. 1616–1625.



- Best, V., Thompson, E. R., Mason, C. R. & Kidd, G. (2013), 'Spatial release from masking as a function of the spectral overlap of competing talkers', *The Journal of the Acoustical Society of America*, 133(6), pp. 3677–3680.
- Blank, M. A. & Foss, D. J. (1978), 'Semantic facilitation and lexical access during sentence processing', *Memory & Cognition*, 6(6), pp. 644–652.
- Blattner, M. M., Sumikawa, D. & Greenberg, R. (1989), 'Earcons and icons: Their structure and common design principles', *Human-Computer Interaction*, 4(1), pp. 11–44.
- Blauert, J. (1997), *Spatial Hearing: The Psychophysics of Human Sound Localization*, The MIT Press, Cambridge, MA, revised edition.
- Blumlein, A. D. (1933), 'Improvements in and Relating to Sound-transmission, Sound-recording and Sound-reproducing Systems', [online], Available: [https://worldwide.espacenet.com/publicationDetails/biblio?CC=GB&NR=394325&KC=&FT=E&locale=en\\_EP](https://worldwide.espacenet.com/publicationDetails/biblio?CC=GB&NR=394325&KC=&FT=E&locale=en_EP), [Accessed: 2016-09-16].
- Blurton, S. P., Greenlee, M. W. & Gondan, M. (2015), 'Cross-modal cueing in audiovisual spatial attention', *Attention, Perception, & Psychophysics*, 77(7), pp. 2356–2376.
- Boersma, P. & Weenink, D. (n.d.), 'Praat: Doing Phonetics by Computer', [online], Available: [www.praat.org](http://www.praat.org), [Accessed: 2015-02-27].
- Bolia, R. S., Nelson, W. T., Ericson, M. A. & Simpson, B. D. (2000), 'A speech corpus for multitalker communications research', *The Journal of the Acoustical Society of America*, 107(2), pp. 1065–1066.
- Bolia, R. S., Nelson, W. T. & Morley, R. M. (2001), 'Asymmetric performance in the cocktail party effect: Implications for the design of spatial audio displays', *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 43(2), pp. 208–216.
- Borodin, Y., Bigham, J. P., Dausch, G. & Ramakrishnan, I. V. (2010), 'More than meets the eye: A survey of screen-reader browsing strategies', in: 'Proceedings of the 2010 International Cross-Disciplinary Conference on Web Accessibility (W4A)', Article no. 13, [online], Available: <http://doi.acm.org/10.1145/1805986.1806005>, [Accessed: 2013-08-30].
- Botte, M.-C. (1995), 'Auditory attentional bandwidth: Effect of level and frequency range', *The Journal of the Acoustical Society of America*, 98(5), pp. 2475–2485.

- Braun, V. & Clarke, V. (2006), 'Using thematic analysis in psychology', *Qualitative Research in Psychology*, 3(2), pp. 77–101.
- Brazil, E. & Fernström, M. (2006), 'Investigating concurrent auditory icon recognition', in: Stockman, T., Nickerson, L., Frauenberger, C., Edwards, A. D. N. & Brock, D. (Eds.), 'Proceedings of the 12th International Conference on Auditory Display (ICAD2006)', pp. 51–58.
- Brazil, E. & Fernström, M. (2011), 'Auditory icons', in: Hermann, T., Hunt, A. & Neuhoff, J. G. (Eds.), 'The Sonification Handbook', chapter 13, pp. 325–338, Logos Publishing House, Berlin, Germany.
- Brazil, E., Fernström, M. & Bowers, J. (2009), 'Exploring concurrent auditory icon recognition', in: 'Proceedings of the 15th International Conference on Auditory Display (ICAD2009)', [online], Available: <http://www.dev.icad.org/Proceedings/2009/BrazilFernstromBowers2009.pdf>, [Accessed: 2012-11-13].
- Bregman, A. S. (1990), *Auditory Scene Analysis: The Perceptual Organization of Sound*, MIT Press, Cambridge, MA.
- Bregman, A. S. & Campbell, J. (1971), 'Primary auditory stream segregation and perception of order in rapid sequences of tones', *Journal of Experimental Psychology*, 89(2), pp. 244–249.
- Brewster, S. A., Capriotti, A. & Hall, C. (1998), 'Using compound earcons to represent hierarchies', *HCI Letters*, 1(1), pp. 6–8.
- Brewster, S. A., Raty, V. P. & Kortekangas, A. (1996), 'Earcons as a method of providing navigational cues in a menu hierarchy', in: 'Proceedings of HCI'96', pp. 167–183.
- Brewster, S. A., Wright, P. C. & Edwards, A. D. N. (1993), 'An evaluation of earcons for use in auditory human-computer interfaces', in: 'Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '93', pp. 222–227.
- Brewster, S. A., Wright, P. C. & Edwards, A. D. N. (1995a), 'Experimentally derived guidelines for the creation of earcons', in: Allen, G., Wilkinson, J. & Wright, P. (Eds.), 'Adjunct Proceedings of HCI'95', pp. 155–159.

- Brewster, S. A., Wright, P. C. & Edwards, A. D. N. (1995b), 'Parallel earcons: Reducing the length of audio messages', *International Journal of Human-Computer Studies*, 43(2), pp. 153–175.
- Brock, D., McClimens, B., Gregory Trafton, J., McCurry, M. & Perzanowski, D. (2008), 'Evaluating listeners' attention to and comprehension of spatialized concurrent and serial talkers at normal and synthetically faster rate of speech', in: 'Proceedings of the 14th International Conference on Auditory Display (ICAD2008)', [online], Available: <http://www.icad.org/node/2336>, [Accessed: 2013-06-03].
- Brokx, J. P. L. & Nootboom, S. G. (1982), 'Intonation and the perceptual separation of simultaneous voices', *Journal of Phonetics*, 10(1), pp. 23–36.
- Brown, A., Aizpurua, A., Jay, C., Evans, M., Glancy, M., Harper, S. (N.D.), 'Contrasting delivery modes for second screen TV content — push or pull?', [Unpublished manuscript provided in personal correspondence (2017)].
- Brown, A., Evans, M., Jay, C., Glancy, M., Jones, R. & Harper, S. (2014), 'HCI over multiple screens', in: 'Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems - CHI EA '14', pp. 665–674.
- Brown, A., Jay, C. & Harper, S. (2012), 'Tailored presentation of dynamic web content for audio browsers', *International Journal of Human-Computer Studies*, 70(3), pp. 179–196.
- Brungart, D. S. (2001), 'Informational and energetic masking effects in the perception of two simultaneous talkers', *The Journal of the Acoustical Society of America*, 109(3), pp. 1101–1109.
- Brungart, D. S., Chang, P. S., Simpson, B. D. & Wang, D. (2009), 'Multitalker speech perception with ideal time-frequency segregation: Effects of voice characteristics and number of talkers', *The Journal of the Acoustical Society of America*, 125(6), pp. 4006–4022.
- Brungart, D. S., Ericson, M. A. & Simpson, B. D. (2002), 'Design considerations for improving the effectiveness of multitalker speech displays', in: Nakatsu, R. & Kawahara, H. (Eds.), 'Proceedings of the 8th International Conference on Auditory Display', [online], Available: <http://www.icad.org/node/2738>, [Accessed: 2013-07-15].

- Brungart, D. S. & Simpson, B. D. (2001), 'Distance-based speech segregation in near-field virtual audio displays', in: 'Proceedings of the 7th International Conference on Auditory Display (ICAD2001)', pp. 169–174.
- Brungart, D. S. & Simpson, B. D. (2002), 'The effects of spatial separation in distance on the informational and energetic masking of a nearby speech signal', *The Journal of the Acoustical Society of America*, 112(2), pp. 664–676.
- Brungart, D. S. & Simpson, B. D. (2003), 'Optimizing the spatial configuration of a seven-talker speech display', in: Brazil, E. & Shinn-Cunningham, B. (Eds.), 'Proceedings of the 9th International Conference on Auditory Display (ICAD2003)', pp. 188–191.
- Brungart, D. S. & Simpson, B. D. (2005a), 'Improving multitalker speech communication with advanced audio displays', in: 'New Directions for Improving Audio Effectiveness. Meeting Proceedings RTO-MP-HFM-123', Article no. 30, [online], Available: <http://oai.dtic.mil/oai/oai?verb=getRecord&metadataPrefix=html&identifier=ADA454531>, [Accessed: 2016-09-24].
- Brungart, D. S. & Simpson, B. D. (2005b), 'Optimizing the spatial configuration of a seven-talker speech display', *ACM Transactions on Applied Perception*, 2(4), pp. 430–436.
- Brungart, D. S. & Simpson, B. D. (2007), 'Effect of target-masker similarity on across-ear interference in a dichotic cocktail-party listening task', *The Journal of the Acoustical Society of America*, 122(3), pp. 1724–1734.
- Brungart, D. S., Simpson, B. D., Ericson, M. A. & Scott, K. R. (2001), 'Informational and energetic masking effects in the perception of multiple simultaneous talkers', *The Journal of the Acoustical Society of America*, 110(5), pp. 2527–2538.
- Butler, J. G. (2007), 'Style and sound', in: Butler, J. G. (Ed.), 'Television: Critical methods and applications', chapter 8, pp. 246–270, Lawrence Erlbaum Associates, Mahwah, NJ, 3rd edition.
- Canter, D., Rivers, R. & Storrs, G. (1985), 'Characterizing user navigation through complex data structures', *Behaviour & Information Technology*, 4(2), pp. 93–102.
- Cao, A., Chintamani, K. K., Pandya, A. K. & Ellis, R. D. (2009), 'NASA TLX: Software for assessing subjective mental workload', *Behavior Research Methods*, 41(1), pp. 113–117.

- Carlander, O., Kindström, M. & Eriksson, L. (2005), 'Intelligibility of stereo and 3D-audio call signs for fire and rescue command operators', in: Brazil, E. (Ed.), 'Proceedings of the 11th International Conference on Auditory Display (ICAD2005)', pp. 292–295.
- Carroll, N. (2003), *Engaging the Moving Image*, Yale University Press, London.
- Cassidy, G. & MacDonald, R. A. (2007), 'The effect of background music and background noise on the task performance of introverts and extraverts', *Psychology of Music*, 35(3), pp. 517–537.
- Cauchard, F., Cane, J. E. & Weger, U. W. (2012), 'Influence of background speech and music in interrupted reading: An eye-tracking study', *Applied Cognitive Psychology*, 26(3), pp. 381–390.
- Chalikia, M. H. & Bregman, A. S. (1989), 'The perceptual segregation of simultaneous auditory signals: Pulse train segregation and vowel segregation', *Perception & Psychophysics*, 46(5), pp. 487–496.
- Chapdelaine, C. (2010), 'In-situ study of blind individuals listening to audio-visual contents', in: 'Proceedings of the 12th International ACM SIGACCESS Conference on Computers and Accessibility', ASSETS '10, pp. 59–66.
- Chapdelaine, C. (2012), 'Specialized DVD player to render audio description and its usability performance', in: 'Proceedings of the 14th International ACM SIGACCESS Conference on Computers and Accessibility - ASSETS '12', pp. 203–204.
- Chapdelaine, C. & Gagnon, L. (2009), 'Accessible videodescription on-demand', in: 'Proceedings of the 11th International ACM SIGACCESS Conference on Computers and Accessibility', Assets '09, pp. 221–222.
- Cherry, E. C. (1953), 'Some experiments on the recognition of speech, with one and two ears', *The Journal of the Acoustical Society of America*, 25(5), pp. 975–979.
- Consumer Electronics Association (2014), 'CEA and NATPE Joint National Survey Reveals Opportunities to Improve Synchronized Program Content', 'Consumer Electronics Association', [online], Available: <https://www.ce.org/News/News-Releases/Press-Releases/2013-Press-Releases/CEA-and-NATPE-Joint-National-Survey-Reveals-Opportunities-to-Improve-Synchronized-Program-Content.aspx?feed=Events-Press-Releases>, [Accessed: 2014-11-26].

- Culling, J. F. & Darwin, C. J. (1993), 'Perceptual separation of simultaneous vowels: Within and across-formant grouping by F0', *The Journal of the Acoustical Society of America*, 93(6), pp. 3454–3467.
- Dai, H., Scharf, B. & Buus, S. (1991), 'Effective attenuation of signals in noise under focused attention', *The Journal of the Acoustical Society of America*, 89(6), pp. 2837–2842.
- Darwin, C. J., Brungart, D. S. & Simpson, B. D. (2003), 'Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers', *The Journal of the Acoustical Society of America*, 114(5), pp. 2913–2922.
- Delgutte, B. (1990), 'Physiological mechanisms of psychophysical masking: Observations from auditory-nerve fibers', *The Journal of the Acoustical Society of America*, 87(2), pp. 791–809.
- DeValois, R. L. & DeValois, K. K. (1990), *Spatial Vision*, Oxford Psychology Series, Oxford University Press, New York.
- Devas, F. (2016), 'Cities', 'Planet Earth II', [Television Programme], Episode: 6, British Broadcasting Corporation.
- Digital UK (2012), 'Digital TV Switchover 2008-2012 Final Report', Technical report, [online], Available: [http://www.digitaluk.co.uk/\\_data/assets/pdf\\_file/0019/82324/DigitalUK\\_Switchoverfinal\\_report\\_Nov2012.pdf](http://www.digitaluk.co.uk/_data/assets/pdf_file/0019/82324/DigitalUK_Switchoverfinal_report_Nov2012.pdf), [Accessed: 2016-05-04].
- Dingler, T., Lindsay, J. & Walker, B. N. (2008), 'Learnability of sound cues for environmental features: Auditory icons, earcons, spearcons and speech', in: 'Proceedings of the International Conference on Auditory Display (ICAD 2008)', [online], Available: <http://sonify.psych.gatech.edu/publications/pdfs/2008ICAD-DinglerLindsayWalker.pdf>, [Accessed: 2012-10-24].
- Dirican, A. C. & Göktürk, M. (2011), 'Psychophysiological measures of human cognitive states applied in human computer interaction', *Procedia Computer Science*, 3, pp. 1361–1367.
- Dix, A., Finlay, J., Abowd, G. D. & Beale, R. (2004), *Human-Computer Interaction*, Pearson Education Limited, Harlow, UK, 3rd edition.

- Dobrovolsky, M. & Katamba, F. (1996), 'Phonetics: The sounds of language', in: O'Grady, W., Dobrovolsky, M. & Katamba, F. (Eds.), 'Contemporary Linguistics', chapter 2, pp. 18–67, Longman, Harlow, UK, 3rd edition.
- Dombois, F. & Eckel, G. (2011), 'Audification', in: Hermann, T., Hunt, A. & Neuhoff, J. G. (Eds.), 'The Sonification Handbook', chapter 12, pp. 301–324, Logos Publishing House, Berlin, Germany.
- Dowell, J., Malacria, S., Kim, H. & Anstead, E. (2015), 'Companion apps for information-rich television programmes: representation and interaction', *Personal and Ubiquitous Computing*, 19(7), pp. 1215–1228.
- Driver, J. (1996), 'Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading', *Nature*, 381(6577), pp. 66–68.
- Driver, J. & Spence, C. J. (1994), 'Spatial synergies between auditory and visual attention', in: Umiltà, C. & Moscovitch, M. (Eds.), 'Attention and Performance XV: Conscious and Nonconscious Information Processing', chapter 12, pp. 311–331, MIT Press, Cambridge, MA.
- Drullman, R. & Bronkhorst, A. W. (2000), 'Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation', *The Journal of the Acoustical Society of America*, 107(4), pp. 2224–2235.
- Durlach, N. I., Mason, C. R., Gallun, F. J., Shinn-Cunningham, B., Colburn, H. S. & Kidd Jr., G. (2005), 'Informational masking for simultaneous nonspeech stimuli: Psychometric functions for fixed and randomly mixed maskers', *The Journal of the Acoustical Society of America*, 118(4), pp. 2482–2497.
- Elsweiler, D., Mandl, S. & Kirkegaard Lunn, B. (2010), 'Understanding casual-leisure information needs: a diary study in the context of television viewing', in: 'Proceedings of the Third Symposium on Information Interaction in Context (IliX)', pp. 25–34.
- Encelle, B., Beldame, M. O. & Prié, Y. (2013), 'Towards the usage of pauses in audio-described videos', in: 'Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility', W4A '13, Article no. 31, [online], Available: <http://doi.acm.org/http://dx.doi.org/10.1145/2461121.2461130>, [Accessed: 2013-12-19].

- Encelle, B., Ollagnier-Beldame, M., Pouchot, S. & Prié, Y. (2011), 'Annotation-based video enrichment for blind people: A pilot study on the use of earcons and speech synthesis', in: 'The Proceedings of the 13th International ACM SIGACCESS Conference on Computers and Accessibility', ASSETS '11, pp. 123–130.
- Erbes, V., Schultz, F., Lindau, A. & Weinzierl, S. (2012), 'An extraaural headphone system for optimized binaural reproduction', in: '38th German Annual Conference on Acoustics', pp. 313–314.
- Ericson, M. A., Brungart, D. S. & Simpson, B. D. (2004), 'Factors that influence intelligibility in multitalker speech displays', *The International Journal of Aviation Psychology*, 14(3), pp. 313–334.
- Ericson, M. A. & McKinley, R. L. (1997), 'The intelligibility of multiple talkers separated spatially in noise', in: Gilkey, R. H. & Anderson, T. R. (Eds.), 'Binaural and Spatial Hearing in Real and Virtual Environments', December 1987, chapter 32, Lawrence Erlbaum Associates, Mahwah, NJ.
- European Blind Union (EBU) (2008), 'Digital TV Accessibility: Report on the Current Status in European Countries', [online], Available: [http://www.euroblind.org/media/eplica/EBU\\_DigitalTV\\_Report\\_2008.doc](http://www.euroblind.org/media/eplica/EBU_DigitalTV_Report_2008.doc), [Accessed: 2016-09-16].
- Fairbanks, G., Guttman, N. & Miron, M. S. (1957a), 'Auditory comprehension of repeated high-speed messages', *The Journal of Speech and Hearing Disorders*, 22(1), pp. 20–22.
- Fairbanks, G., Guttman, N. & Miron, M. S. (1957b), 'Effects of time compression upon the comprehension of connected speech', *The Journal of Speech and Hearing Disorders*, 22(1), pp. 10–19.
- Feddersen, W. E., Sandel, T. T., Teas, D. C. & Jeffress, L. A. (1957), 'Localization of high-frequency tones', *The Journal of the Acoustical Society of America*, 29(9), pp. 988–991.
- Fellowes, J. (2012), 'Autistic spectrum, captions and audio description', *Revista Brasileira de Tradução Visual*, 13(13), [online], Available: <http://www.rbtv.associadosdainclusao.com.br/index.php/principal/article/view/162/275>, [Accessed: 2014-12-11].
- FFmpeg Developers (n.d.), 'FFmpeg', [online], Available: <https://ffmpeg.org/>, [Accessed: 2016-09-16].



- Field, A. (2009), *Discovering Statistics Using SPSS*, Introducing Statistical Methods, SAGE Publications, London, UK, 3rd edition.
- Findlay, J. M. & Gilchrist, I. D. (2003), *Active Vision*, Oxford University Press, Oxford, UK.
- Fogerty, D., Kewley-Port, D. & Humes, L. E. (2012), 'Temporal offset judgments for concurrent vowels by young, middle-aged, and older adults', *The Journal of the Acoustical Society of America Express Letters*, 131(6), pp. EL499–505.
- Forrester, I. (2012), 'What is Perceptive Media?', 'BBC Research & Development Blog', [online], Available: <http://www.bbc.co.uk/blogs/researchanddevelopment/2012/07/what-is-perceptive-media.shtml>, [Accessed: 2016-06-07].
- Frauenberger, C. (2013), 'Personal Correspondence'.
- Frauenberger, C., Putz, V. & Holdrich, R. (2004), 'Spatial auditory displays - A study on the use of virtual audio environments as interfaces for users with visual disabilities', in: 'Proceedings of the 7th International Conference on Digital Audio Effects (DAFx'04)', pp. 384–389.
- Frauenberger, C., Putz, V., Höldrich, R. & Stockman, T. (2005a), 'An auditory 3D file manager designed from interaction patterns', in: 'Proceedings of the 8th International Conference on Digital Audio Effects (DAFx'05)', [online], Available: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.476.7231&rep=rep1&type=pdf>, [Accessed: 2013-05-31].
- Frauenberger, C., Putz, V., Höldrich, R. & Stockman, T. (2005b), 'Interaction patterns for auditory user interfaces', in: 'Proceedings of ICAD 05-Eleventh Meeting of the International Conference on Auditory Display', pp. 154–161.
- Frauenberger, C. & Stockman, T. (2006), 'Patterns in auditory menu design', in: Stockman, T., Nickerson, L. V., Frauenberger, C., Edwards, A. D. N. & Brock, D. (Eds.), 'Proceedings of the 12th International Conference on Auditory Display (ICAD2006)', pp. 141–147.
- Frauenfelder, U. H. & Tyler, L. K. (1987), 'The process of spoken word recognition: An introduction', *Cognition*, 25(1), pp. 1–20.

- Freeview (n.d.), 'Freeview HD Recorder', 'freeview.co.uk', [online], Available: <http://www.freeview.co.uk/what-we-offer/freeview-recorder>, [Accessed: 2015-10-18].
- Freyman, R. L., Balakrishnan, U. & Helfer, K. S. (2001), 'Spatial release from informational masking in speech recognition', *The Journal of the Acoustical Society of America*, 109(5), pp. 2112–2122.
- Freyman, R. L., Helfer, K. S., McCall, D. D. & Clifton, R. K. (1999), 'The role of perceived spatial separation in the unmasking of speech', *The Journal of the Acoustical Society of America*, 106(6), pp. 3578–3588.
- Furnham, A. & Strbac, L. (2002), 'Music is as distracting as noise: The differential distraction of background music and noise on the cognitive test performance of introverts and extraverts', *Ergonomics*, 45(3), pp. 203–217.
- García, S., Fernández, A., Luengo, J. & Herrera, F. (2010), 'Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power', *Information Sciences*, 180(10), pp. 2044–2064.
- Gaver, W. (1986), 'Auditory icons: Using sound in computer interfaces', *Human-Computer Interaction*, 2(2), pp. 167–177.
- Gaver, W. (1989), 'The SonicFinder: An interface that uses auditory icons', *Human-Computer Interaction*, 4(1), pp. 67–94.
- Gaver, W. W., Smith, R. B. & O'Shea, T. (1991), 'Effective sounds in complex situations: The ARKola simulation', in: Robertson, S. P., Olson, G. M. & Olson, J. S. (Eds.), 'Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '91', pp. 85–90.
- Gelfand, S. A. (2010), *Hearing: An Introduction to Psychological and Physiological Acoustics*, Informa Healthcare, London, UK, 5th edition.
- Gillan, J. (2011), *Television and New Media*, Taylor & Francis Ltd, New York, NY, USA.
- Glasberg, B. R. & Moore, B. C. J. (1990), 'Derivation of auditory filter shapes from notched-noise data', *Hearing Research*, 47(1-2), pp. 103–138.
- Gopher, D. (1973), 'Eye-movement patterns in selective listening tasks of focused attention', *Perception & Psychophysics*, 14(2), pp. 259–264.

- Götze, M. & Strothotte, T. (2001), 'An approach to help functionally illiterate people with graphical reading aids', in: '1st International Symposium on Smart Graphics', [online], Available: <http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.14.9227>, [Accessed: 2016-04-30].
- Gradinar, A., Burnett, D., Coulton, P., Forrester, I., Watkins, M., Scutt, T. & Murphy, E. (2015), 'Perceptive media – Adaptive storytelling for digital broadcast', in: Abascal, J., Barbosa, S., Fetter, M., Gros, T., Palanque, P. & Winckler, M. (Eds.), 'Human-Computer Interaction INTERACT 2015', *Lecture Notes in Computer Science*, volume 9299, pp. 586–589, Springer International Publishing.
- Green, K. P. & Kuhl, P. K. (1989), 'The role of visual information in the processing of place and manner features in speech perception', *Perception & Psychophysics*, 45(1), pp. 34–42.
- Green, K. P., Kuhl, P. K., Meltzoff, A. N. & Stevens, E. B. (1991), 'Integrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the McGurk effect', *Perception & Psychophysics*, 50(6), pp. 524–536.
- Greenberg, G. Z. & Larkin, W. D. (1968), 'Frequency-response characteristic of auditory observers detecting signals of a single frequency in noise: The probe-signal method', *The Journal of the Acoustical Society of America*, 44(6), pp. 1513–1523.
- Gribbons, B. (2008), 'Universal accessibility and functionally illiterate populations', in: Jacko, J. A. & Sears, A. (Eds.), 'The Human-Computer Interaction Guidebook: Fundamentals, Evolving Technologies and Emerging Applications', chapter 44, pp. 872–881, Lawrence Erlbaum Associates, Mahwah, NJ, 2nd edition.
- Grond, F. & Berger, J. (2011), 'Parameter mapping sonification', in: Hermann, T., Hunt, A. & Neuhoff, J. G. (Eds.), 'The Sonification Handbook', chapter 15, pp. 363–397, Logos Publishing House, Berlin, Germany.
- Guerreiro, J. (2013), 'Using simultaneous audio sources to speed-up blind people's web scanning', in: 'Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility', Article no. 8, [online], Available: <http://dl.acm.org/citation.cfm?id=2461121.2461154>, [Accessed: 2013-08-01].
- Guerreiro, J. & Gonçalves, D. (2014), 'Text-To-Speeches: Evaluating the perception of concurrent speech by blind people', in: 'Proceedings of ASSETS'14', pp. 169–176.

- Guerreiro, J. & Gonçalves, D. (2016), 'Scanning for digital content: How blind and sighted people perceive concurrent speech', *ACM Transactions on Accessible Computing*, 8(1), pp. 1–28.
- Hand, J. (2012), 'Ceefax Service Closes Down After 38 Years on BBC', 'BBC News', [online], Available: <http://www.bbc.co.uk/news/uk-20032882>, [Accessed: 2016-06-02].
- Harley, T. A. (2014), *The Psychology of Language*, Psychology Press, Hove, UK, 4th edition.
- Harrington, S., Highfield, T. & Bruns, A. (2013), 'More than a backchannel: Twitter and television', *Participations: Journal of Audience and Reception Studies*, 10(1), pp. 405–409.
- Hart, S. G. (2006), 'NASA-Task Load Index (NASA-TLX); 20 Years Later', in: 'Proceedings of the Human Factors and Ergonomics Society 50th Annual Meeting', pp. 904-908.
- Hart, S. G. & Staveland, L. E. (1988), 'Development of the NASA-TLX (Task Load Index): Results of empirical and theoretical work', in: Hancock, P. & Meshkati, N. (Eds.), 'Human Mental Workload', North Holland Press, Amsterdam, The Netherlands.
- Hawley, M. L., Litovsky, R. Y. & Colburn, H. S. (1999), 'Speech intelligibility and localization in a multi-source environment', *The Journal of the Acoustical Society of America*, 105(6), pp. 3436–3448.
- Hayward, C. (1994), 'Listening to the earth sing', in: Kramer, G. (Ed.), 'Auditory Display: Sonification, Audification, and Auditory Interfaces', pp. 369–404, Addison-Wesley Publishing Company, Reading, MA.
- HbbTV Association (2015), 'HbbTV 2.0 Specification', [online], Available: [https://www.hbbtv.org/wp-content/uploads/2015/07/HbbTV\\_specification\\_2.0.pdf](https://www.hbbtv.org/wp-content/uploads/2015/07/HbbTV_specification_2.0.pdf), [Accessed: 2016-07-20].
- Hecht, J. (2014), 'Netflix Chief Downplays Nielsen Plans to Measure Streaming Service Viewership', 'The Hollywood Reporter', [online], Available: <http://www.hollywoodreporter.com/news/netflix-chief-downplays-nielsen-plans-751931>, [Accessed: 2016-08-14].
- Hedrick, M. S. & Madix, S. G. (2009), 'Effect of vowel identity and onset asynchrony on concurrent vowel identification', *Journal of Speech, Language, and Hearing Research*, 52(3), pp. 696–705.

- Heiman, G. W., Leo, R. J., Leighbody, G. & Bowler, K. (1986), 'Word intelligibility decrements and the comprehension of time-compressed speech', *Perception & Psychophysics*, 40(6), pp. 407–411.
- Helfer, K. S. (1997), 'Auditory and auditory-visual perception of clear and conversational speech', *Journal of Speech Language and Hearing Research*, 40(2), pp. 432–443.
- Helfer, K. S. & Freyman, R. L. (2009), 'Lexical and indexical cues in masking by competing speech', *The Journal of the Acoustical Society of America*, 125(1), pp. 447–456.
- Henning, B. (1980), 'Some observations on the lateralization of complex waveforms', *The Journal of the Acoustical Society of America*, 68(2), pp. 446–454.
- Hermann, T. (2008), 'Taxonomy and definitions for sonification and auditory display', in: 'Proceedings of the 14th International Conference on Auditory Display (ICAD2008)', [online], Available: <http://hdl.handle.net/1853/49960>, [Accessed: 2016-02-12].
- Hermann, T. (2011), 'Model-based sonification', in: Hermann, T., Hunt, A. & Neuhoff, J. G. (Eds.), 'The Sonification Handbook', chapter 16, pp. 399–427, Logos Publishing House, Berlin, Germany.
- Hermann, T., Meinicke, P., Bekel, H., Ritter, H., Müller, H. M. & Weiss, S. (2002), 'Sonifications For EEG data analysis', in: 'Proceedings of the 2002 International Conference on Auditory Display', [online], Available: <http://hdl.handle.net/1853/51378>, [Accessed: 2016-07-27].
- Hermann, T. & Ritter, H. (1999), 'Listen to your data: Model-based sonification for data analysis', in: Lasker, G. E. & Syed, M. R. (Eds.), 'Advances in Intelligent Computing and Multimedia Systems', pp. 189–194.
- Hitchcock, A. (1960), 'Psycho', [Film], Paramount Pictures.
- Hoare, C. & Hinde, A. F. (2016), 'Television and additional media activity: A taxonomy', Technical report, [online], Available: [https://figshare.com/articles/Television\\_and\\_additional\\_media\\_activity\\_A\\_taxonomy/3856164](https://figshare.com/articles/Television_and_additional_media_activity_A_taxonomy/3856164), [Accessed: 2016-09-24].
- Hochberg, Y. (1988), 'A sharper Bonferroni procedure for multiple tests of significance', *Biometrika*, 75(4), pp. 800–802.
- Hofman, P. M., Van Riswick, J. G. & Opstal, A. J. V. (1998), 'Relearning sound localization with new ears', *Nature Neuroscience*, 1(5), pp. 417–421.

- Holman, T. (2010), *Sound for Film and Television*, Focal Press/Elsevier, Burlington, MA, 3rd edition.
- Holmes, M. E., Josephson, S. & Carney, R. E. (2012), 'Visual attention to television programs with a second-screen application', in: 'Proceedings of the Symposium on Eye Tracking Research and Applications - ETRA '12', pp. 397–400.
- House, A. S., Williams, C. E., Hecker, M. H. L. & Kryter, K. D. (1965), 'Articulation-testing methods: Consonantal differentiation with a closed-response set', *The Journal of the Acoustical Society of America*, 37(1), pp. 158–166.
- Howard, I. P. & Templeton, W. B. (1966), *Human Spatial Orientation*, John Wiley & Sons, London, UK.
- Huenerfauth, M. P. (2002), *Developing Design Recommendations for Computer Interfaces Accessible to Illiterate Users*, Master's dissertation, National University of Ireland, [online], Available: [eniac.cs.qc.cuny.edu/matt/pubs/huenerfauth-2002-thesis.pdf](http://eniac.cs.qc.cuny.edu/matt/pubs/huenerfauth-2002-thesis.pdf), [Accessed: 2016-05-03].
- IBM (n.d.), 'IBM Knowledge Center', IBM, [online], Available: [http://www-01.ibm.com/support/knowledgecenter/SSLVMB\\_21.0.0/com.ibm.spss.statistics\\_21.kc.doc/pv\\_welcome.html](http://www-01.ibm.com/support/knowledgecenter/SSLVMB_21.0.0/com.ibm.spss.statistics_21.kc.doc/pv_welcome.html), [Accessed: 2015-02-27].
- Ibrahim, K. F. & Trundle, E. (2007), *Newnes Guide to Television and Video Technology*, Newnes, Oxford, UK.
- Ihlefeld, A. & Shinn-Cunningham, B. (2008a), 'Spatial release from energetic and informational masking in a divided speech identification task', *The Journal of the Acoustical Society of America*, 123(6), pp. 4380–4392.
- Ihlefeld, A. & Shinn-Cunningham, B. G. (2008b), 'Spatial release from energetic and informational masking in a selective speech identification task', *The Journal of the Acoustical Society of America*, 123(6), pp. 4369–4379.
- Ikei, Y., Yamazaki, H., Hirota, K. & Hirose, M. (2006), 'vCocktail: Multiplexed-voice menu presentation method for wearable computers', in: 'Proceedings of the IEEE Virtual Reality Conference', pp. 183–190.

- Incorporated Television Company (ITC) (2000), ‘ITC Guidance on Standards for Audio Description’, Technical report, Ofcom, [online], Available: [http://stakeholders.ofcom.org.uk/broadcasting/guidance/other-guidance/tv\\_access\\_serv/archive/audio\\_description\\_standards/](http://stakeholders.ofcom.org.uk/broadcasting/guidance/other-guidance/tv_access_serv/archive/audio_description_standards/), [Accessed: 2013-12-19].
- ITU-R (2011), ‘Recommendation ITU-R BS.1770-2: Algorithms to Measure Audio Programme Loudness and True-peak Audio Level’, Technical report, International Telecommunication Union, [online], Available: [https://www.itu.int/dms\\_pubrec/itu-r/rec/bs/R-REC-BS.1770-2-201103-S!!PDF-E.pdf](https://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.1770-2-201103-S!!PDF-E.pdf), [Accessed: 2016-09-08].
- ITU-R (2012), ‘Recommendation ITU-R BS.775-3: Multichannel Stereophonic Sound System With and Without Accompanying Picture’, Technical report, International Telecommunication Union, [online], Available: [https://www.itu.int/dms\\_pubrec/itu-r/rec/bs/R-REC-BS.775-3-201208-I!!PDF-E.pdf](https://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.775-3-201208-I!!PDF-E.pdf), [Accessed: 2015-12-10].
- Iyer, N. & Brungart, D. S. (2010), ‘The effect of disrupting intonation patterns on the perceptual segregation of competing voices’, *The Journal of the Acoustical Society of America*, 128(4), p. 2320.
- Janata, P. & Childs, E. (2004), ‘Marketbuzz: Sonification of real-time financial data’, in: Barrass, S. & Vickers, P. (Eds.), ‘Proceedings of ICAD 04. Tenth Meeting of the International Conference on Auditory Display’, [online], Available: <http://hdl.handle.net/1853/50899>, [Accessed: 2016-07-27].
- Jaye, V. (2012), ‘Making Great TV Even Better: The BBC’s Approach to Companion Experiences’, ‘BBC Internet Blog’, [online], Available: [http://www.bbc.co.uk/blogs/bbcinternet/2012/05/making\\_great\\_tv\\_even\\_better\\_th.html](http://www.bbc.co.uk/blogs/bbcinternet/2012/05/making_great_tv_even_better_th.html), [Accessed: 2013-12-10].
- Jensen, J. F. (2008), ‘Interactive television - A brief media history’, in: Tscheligi, M., Obrist, M. & Lugmayr, A. (Eds.), ‘Changing Television Environments’, *Lecture Notes in Computer Science*, volume 5066, pp. 1–10, Springer Berlin Heidelberg.
- Jeon, M. & Walker, B. N. (2009), ‘“Spindex”: Accelerated initial speech sounds improve navigation performance in auditory menus’, in: ‘Proceedings of the Human Factors and Ergonomics Society Annual Meeting’, volume 53, pp. 1081–1085.
- Jeon, M. & Walker, B. N. (2011), ‘Spindex (speech index) improves auditory menu acceptance and navigation performance’, *ACM Transactions on Accessible Computing*, 3(3), pp. 10:1–26.

- Jobst, M. (2012), 'A Faraway Land About Which We Know Nothing', 'Upstairs Downstairs', [Television Programme], Series: 2, Episode: 1, British Broadcasting Corporation.
- Jolly, S. J. E. & Evans, M. J. (2013), 'Improving the experience of media in the connected home with a new approach to inter-device communication', British Broadcasting Corporation, [online], Available: <http://www.bbc.co.uk/rd/publications/whitepaper242>, [Accessed: 2013-11-18].
- Jones, D., Madden, C. & Miles, C. (1992), 'Privileged access by irrelevant speech to short-term memory: The role of changing state', *The Quarterly Journal of Experimental Psychology Section A*, 44(4), pp. 645–669.
- Jones, T. (2011), 'Designing for second screens: The Autumnwatch Companion', 'BBC Research & Development Blog', [online], Available: <http://www.bbc.co.uk/blogs/researchanddevelopment/2011/04/the-autumnwatch-companion---de.shtml>, [Accessed: 2015-10-16].
- Joshi, M., Iyer, M., Gupta, N. & Barreto, A. (2010), 'Effect of gender and sound spatialization on speech intelligibility in multiple speaker environment', in: Sobh, T. & Elleithy, K. (Eds.), 'Innovations in Computing Sciences and Software Engineering', pp. 547–550, Springer Netherlands, Dordrecht.
- Kidd Jr., G., Arbogast, T. L., Mason, C. R. & Gallun, F. J. (2005), 'The advantage of knowing where to listen', *The Journal of the Acoustical Society of America*, 118(6), pp. 3804–3815.
- Kidd Jr., G., Mason, C. R. & Arbogast, T. L. (2002), 'Similarity, uncertainty, and masking in the identification of nonspeech auditory patterns', *The Journal of the Acoustical Society of America*, 111(3), pp. 1367–1376.
- Kimura, D. (1961a), 'Cerebral dominance and the perception of verbal stimuli', *Canadian Journal of Psychology*, 15(3), pp. 166–171.
- Kimura, D. (1961b), 'Some effects of temporal-lobe damage on auditory perception', *Canadian Journal of Psychology*, 15(3), pp. 156–165.
- Kindström, M., Carlander, O. & Eriksson, L. (2006), 'Comparison of audio systems for intelligibility in multitalker speech displays', *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 50(3), pp. 309–313.



- Knecht, S., Dräger, B., Deppe, M., Bobe, L., Lohmann, H., Flöel, A., Ringelstein, E.-B. & Henningsen, H. (2000), 'Handedness and hemispheric language dominance in healthy humans', *Brain: A Journal of Neurology*, 123(12), pp. 2512–2518.
- Knoche, H. & Huang, J. (2012), 'Text is not the enemy: How illiterates' use their mobile phones', in: 'NUIs for New Worlds: New Interaction Forms and Interfaces for Mobile Applications in Developing Countries Workshop at CHI'2012', [online], Available: [https://www.researchgate.net/profile/Hendrik\\_Knoche/publication/266595198\\_Text\\_is\\_not\\_the\\_enemy\\_How\\_illiterates'\\_use\\_their\\_mobile\\_phones/links/5476ebcb0cf245eb43728d4c.pdf](https://www.researchgate.net/profile/Hendrik_Knoche/publication/266595198_Text_is_not_the_enemy_How_illiterates'_use_their_mobile_phones/links/5476ebcb0cf245eb43728d4c.pdf), [Accessed: 2016-04-30].
- Kobayashi, M. & Schmandt, C. (1997), 'Dynamic soundscape', in: 'Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '97', pp. 194–201.
- Koffka, K. (1922), 'Perception: An introduction to the Gestalt-theorie', *The Psychological Bulletin*, 19(10), pp. 531–585.
- Kramer, G. (1994), 'An introduction to auditory display', in: Kramer, G. (Ed.), 'Auditory Display: Sonification, Audification, and Auditory Interfaces', chapter 1, pp. 1–78, Addison-Wesley, Reading, MA.
- Krejtz, I., Szarkowska, A., Krejtz, K., Walczak, A. & Duchowski, A. (2012), 'Audio description as an aural guide of children's visual attention', in: 'Proceedings of the Symposium on Eye Tracking Research and Applications - ETRA '12', pp. 99–106.
- Krejtz, K., Krejtz, I., Duchowski, A., Szarkowska, A. & Walczak, A. (2012), 'Multimodal learning with audio description', in: 'Proceedings of the ACM Symposium on Applied Perception - SAP '12', pp. 83–90.
- LeCompte, D. C. (1994), 'Extending the irrelevant speech effect beyond serial recall', *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 20(6), pp. 1396–1408.
- Lee, J. H. & Humes, L. E. (2012), 'Effect of fundamental-frequency and sentence-onset differences on speech-identification performance of young and older adults in a competing-talker background', *The Journal of the Acoustical Society of America*, 132(3), pp. 1700–1717.
- LG USA (n.d.), 'LG webOS TVs', [online], Available: <http://www.lg.com/us/experience-tvs/smart-tv/use>, [Accessed: 2016-08-19].

- Lin, C.-L., Hung, Y.-H., Chen, H.-Y. & Chu, S.-L. (2012), 'Content-aware smart remote control for Android-based TV', in: '2012 IEEE International Conference on Consumer Electronics (ICCE)', pp. 678–679.
- Lodge, N. K. & Slater, J. N. (1992), 'Helping blind people to watch television - The AUDETEL project', *International Broadcasting Convention*, pp. 86–91.
- Lorho, G., Hiipakka, J. & Marila, J. (2002), 'Structured menu presentation using spatial sound separation', *Human Computer Interaction with Mobile Devices Lecture Notes in Computer Science*, 2411, pp. 419–424.
- Luyten, K., Thys, K., Huypens, S. & Coninx, K. (2006), 'Telebuddies on the move: Social stitching to enhance the networked gaming experience', in: 'Proceedings of 5th ACM SIGCOMM Workshop on Network and System Support for Games', NetGames '06, Article no. 18, [online], Available: <http://doi.acm.org/10.1145/1230040.1230084>, [Accessed: 2013-12-16].
- Marston, D. (2016), 'upmixer\_x', [Software provided in personal correspondence].
- Martin, R. C., Wogalter, M. S. & Forlano, J. G. (1988), 'Reading comprehension in the presence of unattended speech and music', *Journal of Memory and Language*, 27(4), pp. 382–398.
- Martin, R. L., McAnally, K. I., Bolia, R. S., Eberle, G. & Brungart, D. S. (2012), 'Spatial release from speech-on-speech masking in the median sagittal plane', *The Journal of the Acoustical Society of America*, 131(1), pp. 378–385.
- Massaro, D. W. (1970), 'Preperceptual auditory images', *Journal of Experimental Psychology*, 85(3), pp. 411–417.
- Massaro, D. W. (1974), 'Perceptual units in speech recognition', *Journal of Experimental Psychology*, 102(2), pp. 199–208.
- Massaro, D. W. & Idson, W. L. (1977), 'Backward recognition masking in relative pitch judgments', *Perceptual and Motor Skills*, 45(1), pp. 87–97.
- Matt (Samsung) (2012), 'A Galaxy of Remote Possibilities', [online], Available: <http://www.samsung.com/uk/discover/tv/turn-your-tv-on-to-a-galaxy-of-remote-possibilities/>, [Accessed: 2016-08-19].

- McAnally, K. I., Bolia, R. S., Martin, R. L., Eberle, G. & Brungart, D. S. (2002), 'Segregation of multiple talkers in the vertical plane: Implications for the design of a multiple talker display', *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 46(4), pp. 588–591.
- McDowd, J. M. & Filion, D. L. (1992), 'Aging, selective attention, and inhibitory processes: A psychophysiological approach', *Psychology and Aging*, 7(1), pp. 65–71.
- McFadden, D. & Pasanen, E. G. (1976), 'Lateralization at high frequencies based on interaural time differences', *The Journal of the Acoustical Society of America*, 59(3), pp. 634–639.
- McGee-Lennon, M., Wolters, M., McLachlan, R., Brewster, S. A. & Hall, C. (2011), 'Name that tune: Musicons as reminders in the home', in: 'Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems - CHI '11', pp. 2803–2806.
- McGookin, D. & Brewster, S. A. (2011), 'Earcons', in: Hermann, T., Hunt, A. & Neuhoff, J. G. (Eds.), 'The Sonification Handbook', chapter 14, pp. 339–361, Logos Publishing House, Berlin, Germany.
- McGookin, D. K. (2004), *Understanding and Improving the Identification of Concurrently Presented Earcons*, Doctoral thesis, University of Glasgow, [online], Available: [http://dcs.gla.ac.uk/~stephen/papers/theses/david\\_mcgookin\\_thesis.pdf](http://dcs.gla.ac.uk/~stephen/papers/theses/david_mcgookin_thesis.pdf), [Accessed: 2016-09-24].
- McGookin, D. K. & Brewster, S. A. (2001), 'FISHEARS — The design of a multimodal focus and context system', in: 'Vol II of Proceedings of the IHM HCI', [online], Available: [http://ftp.dcs.glasgow.ac.uk/~stephen/papers/IHMHCI2001\\_david.pdf](http://ftp.dcs.glasgow.ac.uk/~stephen/papers/IHMHCI2001_david.pdf), [Accessed: 2012-10-31].
- McGookin, D. K. & Brewster, S. A. (2002), 'Dolphin: The design and initial evaluation of multimodal focus and context', in: 'Proceedings of the 2002 International Conference on Auditory Display', [online], Available: <http://eprints.gla.ac.uk/3193/>, [Accessed: 2013-04-23].
- McGookin, D. K. & Brewster, S. A. (2003), 'An investigation into the identification of concurrently presented earcons', in: Brazil, E. & Shinn-Cunningham, B. (Eds.), 'Proceedings of the 9th International Conference on Auditory Display (ICAD2003)', pp. 42–46.

- McGookin, D. K. & Brewster, S. A. (2004a), 'Space, the final frontier: The identification of concurrently presented earcons in a synthetic spatialized auditory environment', in: Barrass, S. & Vickers, P. (Eds.), 'The 10th Meeting of the International Conference on Auditory Display', [online], Available: <http://www.icad.org/websiteV2.0/Conferences/ICAD2004/papers/mcgookin.brewster.pdf>, [Accessed: 2012-10-31].
- McGookin, D. K. & Brewster, S. A. (2004b), 'Understanding concurrent earcons: Applying auditory scene analysis principles to concurrent earcon recognition', *ACM Transactions on Applied Perception*, 1(2), pp. 130–155.
- McGookin, D. K. & Brewster, S. A. (2006), 'Advantages and issues with concurrent audio presentation as part of an auditory display', in: Stockman, T., Nickerson, L. V., Frauenberger, C., Edwards, A. D. N. & Brock, D. (Eds.), 'Proceedings of the 12th International Conference on Auditory Display (ICAD2006)', [online], Available: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.62.1008&rep=rep1&type=pdf>, [Accessed: 2012-10-31].
- McGurk, H. & MacDonald, J. (1976), 'Hearing lips and seeing voices', *Nature*, 264(5588), pp. 746–748.
- McLachlan, R., McGee-Lennon, M. & Brewster, S. A. (2012), 'The sound of musicians: Investigating the design of musically derived audio cues', in: 'Proceedings of the 18th International Conference on Auditory Display', pp. 148–155.
- Medhi, I., Prasad, A. & Toyama, K. (2007a), 'Optimal audio-visual representations for illiterate users of computers', in: 'Proceedings of the 16th International Conference on World Wide Web - WWW '07', pp. 873–882.
- Medhi, I., Sagar, A. & Toyama, K. (2007b), 'Text-free user interfaces for illiterate and semiliterate users', *Information Technologies and International Development*, 4(1), pp. 37–50.
- Mesgarani, N., Shamma, S., Grant, K. W. & Duraiswami, R. (2003), 'Augmented intelligibility in simultaneous multi-talker environments', in: 'Proceedings of the 2003 International Conference on Auditory Display', pp. 71–74.
- Meyer, D. E. & Schvaneveldt, R. W. (1971), 'Facilitation in recognizing pairs of words: evidence of a dependence between retrieval operations', *Journal of Experimental Psychology*, 90(2), pp. 227–234.

- Micheyl, C. (2006), 'Informational masking: An experimentalist's position statement', in: 'Computational and Systems Neuroscience Workshop: Difficult issues in auditory scene analysis', [online], Available: [http://www.isr.umd.edu/Labs/NSL/Cosyne/masking\\_counterpoint.htm](http://www.isr.umd.edu/Labs/NSL/Cosyne/masking_counterpoint.htm), [Accessed: 2013-02-12].
- Microsoft (2016), 'Cortana on your Windows Phone', [online], Available: <https://support.microsoft.com/en-us/help/11694/windows-phone-cortana-on-your-windows-phone>, [Accessed: 2016-09-17].
- Middelweerd, M. J. & Plomp, R. (1987), 'The effect of speechreading on the speech-reception threshold of sentences in noise', *The Journal of the Acoustical Society of America*, 82(6), pp. 2145–2147.
- Middlebrooks, J. C. & Green, D. M. (1990), 'Directional dependence of interaural envelope delays', *The Journal of the Acoustical Society of America*, 87(5), pp. 2149–2162.
- Miller, G. A. (1947), 'The masking of speech', *Psychological Bulletin*, 44(2), pp. 105–129.
- Mondor, T. A. & Bregman, A. S. (1994), 'Allocating attention to frequency regions', *Perception & Psychophysics*, 56(3), pp. 268–276.
- Mondor, T. A. & Zatorre, R. J. (1995), 'Shifting and focusing auditory spatial attention', *Journal of Experimental Psychology: Human Perception and Performance*, 21(2), pp. 387–409.
- Montpetit, M.-J., Klym, N. & Mirlacher, T. (2011), 'The future of IPTV', *Multimedia Tools and Applications*, 53(3), pp. 519–532.
- Moon, C., Lagercrantz, H. & Kuhl, P. K. (2013), 'Language experienced in utero affects vowel perception after birth: a two-country study', *Acta Paediatrica*, 102(2), pp. 156–160.
- Moore, B. C. J. (2012), *An Introduction to the Psychology of Hearing*, Emerald Group Publishing Limited, Bingley, UK, sixth edition.
- Moore, B. C. J. & Vickers, D. A. (1997), 'The role of spread excitation and suppression in simultaneous masking', *The Journal of the Acoustical Society of America*, 102(4), pp. 2284–2290.
- Moray, N. (1959), 'Attention in dichotic listening: Affective cues and the influence of instructions', *Quarterly Journal of Experimental Psychology*, 11(1), pp. 56–60.

- Mowbray, G. H. (1953), 'Simultaneous vision and audition: The comprehension of prose passages with varying levels of difficulty', *Journal of Experimental Psychology*, 46(5), pp. 365–372.
- Mowbray, G. H. (1954), 'The perception of short phrases presented simultaneously for visual & auditory reception', *Quarterly Journal of Experimental Psychology*, 6(2), pp. 86–92.
- Mullins, A. T. (1996), *Audiostreamer: Leveraging The Cocktail Party Effect for Efficient Listening*, Masters thesis, Massachusetts Institute of Technology, [online], Available: <http://hdl.handle.net/1721.1/34328>, [Accessed: 2013-05-29].
- Mynatt, E. D. (1994a), 'Designing with auditory icons', in: Kramer, G. & Smith, S. (Eds.), 'Proceedings of the 2nd International Conference on Auditory Display (ICAD94)', pp. 109–120.
- Mynatt, E. D. (1994b), 'Designing with auditory icons: How well do we identify auditory cues?', in: 'Association for Computer Machinery Computer Human Interaction Conference (CHI'94)', pp. 269–270.
- Mynatt, E. D., Back, M., Want, R., Baer, M. & Ellis, J. B. (1998), 'Designing Audio Aura', in: 'CHI '98 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems', pp. 566–573.
- Nandakumar, A. & Murray, J. (2014), 'Companion apps for long arc TV series: Supporting new viewers in complex storyworlds with tightly synchronized context-sensitive annotations', in: 'Proceedings of the 2014 ACM International Conference on Interactive Experiences for TV and Online Video - TVX '14', pp. 3–10.
- NASA (n.d.), 'NASA TLX Paper and Pencil Version', [online], Available: <http://humansystems.arc.nasa.gov/groups/tlx/downloads/TLXScale.pdf>, [Accessed: 2015-12-07].
- NASA AMES Research Center (n.d.), 'NASA Task Load Index (TLX) Paper and Pencil Version', Technical report, [online], Available: <http://humansystems.arc.nasa.gov/groups/TLX/downloads/TLX.pdf>, [Accessed: 2015-09-17].
- Neate, T., Evans, M. & Jones, M. (2016), 'Designing visual complexity for dual-screen media', in: 'Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI '16', pp. 475–486.

- Neate, T., Jones, M. & Evans, M. (2015), 'Mediating attention for second screen companion content', in: 'Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15', pp. 3103–3106.
- Neff, D. L., Dethlefs, T. M. & Jesteadt, W. (1993), 'Informational masking for multicomponent maskers with spectral gaps', *The Journal of the Acoustical Society of America*, 94(6), pp. 3112–3126.
- Neff, D. L. & Green, D. M. (1987), 'Masking produced by spectral uncertainty with multicomponent maskers', *Perception & Psychophysics*, 41(5), pp. 409–15.
- Nelson, W. T., Bolia, R. S., Ericson, M. A. & Mckinley, R. L. (1998), 'Monitoring the simultaneous presentation of spatialized speech signals in a virtual acoustic environment', in: 'Proceedings of the 1998 IMAGE Conference', pp. 159–166.
- Nelson, W. T., Bolia, R. S., Ericson, M. A. & McKinley, R. L. (1999), 'Spatial audio displays for speech communications: A comparison of free field and virtual acoustic environments', *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 43(22), pp. 1202–1205.
- Noble, W. & Perrett, S. (2002), 'Hearing speech against spatially separate competing speech versus competing noise', *Perception & Psychophysics*, 64(8), pp. 1325–1336.
- Node.js Foundation (n.d.), 'Node.js', [online], Available: <https://nodejs.org/en/>, [Accessed: 2015-11-28].
- Noland, K. & Truong, L. (2015), 'A Survey of UK Television Viewing Conditions', Technical report, BBC Research & Development, [online], Available: <http://www.bbc.co.uk/rd/publications/whitepaper287>, [Accessed: 2015-10-29].
- Norman, K. L. (1991), *The Psychology of Menu Selection: Designing Cognitive Control at the Human/Computer Interface*, Ablex Pub. Corp., Norwood, N.J.
- Norman, K. L. (2008), 'Better design of menu selection systems through cognitive psychology and human factors', *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 50(3), pp. 556–559.
- Ofcom (2013), 'The Communications Market Report', Technical report, [online], Available: [http://stakeholders.ofcom.org.uk/binaries/research/cmr/cmr13/2013\\_UK\\_CMR.pdf](http://stakeholders.ofcom.org.uk/binaries/research/cmr/cmr13/2013_UK_CMR.pdf), [Accessed: 2016-08-14].

- Ofcom (2015), 'The Communications Market Report', Technical report, [online], Available: [http://stakeholders.ofcom.org.uk/binaries/research/cmr/cmr15/CMR\\_UK\\_2015.pdf](http://stakeholders.ofcom.org.uk/binaries/research/cmr/cmr15/CMR_UK_2015.pdf), [Accessed: 2016-05-18].
- Ofcom (2016), 'The Communications Market Report', Technical report, [online], Available: [http://stakeholders.ofcom.org.uk/binaries/research/cmr/cmr16/uk/CMR\\_UK\\_2016.pdf](http://stakeholders.ofcom.org.uk/binaries/research/cmr/cmr16/uk/CMR_UK_2016.pdf), [Accessed: 2016-08-14].
- Oldfield, R. (1971), 'The assessment and analysis of handedness: The Edinburgh inventory', *Neuropsychologia*, 9(1), pp. 97–113.
- O'Leary, A. & Rhodes, G. (1984), 'Cross-modal effects on visual and auditory object perception', *Perception & Psychophysics*, 35(6), pp. 565–569.
- Oswald, C. J., Tremblay, S. & Jones, D. M. (2000), 'Disruption of comprehension by the meaning of irrelevant sound', *Memory*, 8(5), pp. 345–350.
- Oxenham, A. J. & Moore, B. C. J. (1994), 'Modeling the additivity of nonsimultaneous masking', *Hearing Research*, 80, pp. 105–118.
- Palladino, D. K. (2007), *Efficiency of Spearcon-Enhanced Navigation of One Dimensional Electronic Menus*, Undergraduate thesis, Georgia Institute of Technology, [online], Available: <http://smartech.gatech.edu/handle/1853/26302>, [Accessed: 2012-11-26].
- Palladino, D. K. & Walker, B. N. (2007), 'Learning rates for auditory menus enhanced with spearcons versus earcons', in: Scavone, G. P. (Ed.), 'Proceedings of the 13th International Conference on Auditory Display (ICAD2007)', pp. 274–279.
- Palladino, D. K. & Walker, B. N. (2008a), 'Efficiency of spearcon-enhanced navigation of one dimensional electronic menus', in: 'Proceedings of the 14th International Conference on Auditory Display (ICAD2008)', [online], Available: <http://hdl.handle.net/1853/49908>, [Accessed: 2012-11-29].
- Palladino, D. K. & Walker, B. N. (2008b), 'Navigation efficiency of two dimensional auditory menus using spearcon enhancements', in: 'Proceedings of the Annual Meeting of the Human Factors and Ergonomics Society (HFES2008)', pp. 1262–1266.



- Parente, P. (2008), *Clique: Perceptually Based, Task Oriented Auditory Display for GUI Applications*, Phd thesis, University of North Carolina, [online], Available: <https://www.cs.unc.edu/cms/publications/dissertations/parente.pdf>, [Accessed: 2013-02-01].
- Parseihian, G. & Katz, B. F. G. (2012), 'Morphocons: A new sonification concept based on morphological earcons', *The Journal of the Audio Engineering Society*, 60(6), pp. 409–418.
- Peres, S. C., Best, V., Brock, D., Shinn-Cunningham, B. G., Frauenberger, C., Hermann, T., Neuhoff, J., Nickerson, L. & Stockman, T. (2008), 'Auditory interfaces', in: Kortum, P. (Ed.), 'HCI Beyond the GUI: The Human Factors of Nontraditional Interfaces', chapter 5, pp. 147–196, Morgan Kaufman, San Francisco, CA.
- Perham, N. & Currie, H. (2014), 'Does listening to preferred music improve reading comprehension performance?', *Applied Cognitive Psychology*, 28(2), pp. 279–284.
- Perrott, D. R. & Saberi, K. (1990), 'Minimum audible angle thresholds for sources varying in both elevation and azimuth', *Journal of the Acoustical Society of America*, 87(4), pp. 1728–1731.
- Pettitt, B., Sharpe, K. & Cooper, S. (1996), 'AUDETEL: Enhancing television for visually impaired people', *The British Journal of Visual Impairment*, 14(2), pp. 48–52.
- Pike, C. & Melchior, F. (2013), 'An assessment of virtual surround sound systems for headphone listening of 5.1 multichannel audio', in: 'Audio Engineering Society Convention 134', Article no. 8819, [online], Available: <http://www.aes.org/e-lib/browse.cfm?elib=16720>, [Accessed: 2014-05-09].
- Pitt, I. J. & Edwards, A. D. N. (1997), 'An improved auditory interface for the exploration of lists', in: 'Proceedings of the Fifth ACM International Conference on Multimedia - MULTIMEDIA '97', pp. 51–61.
- Plack, C. J. (2014), *The Sense of Hearing*, Psychology Press, London, UK, 2nd edition.
- Posner, M. I. (1980), 'Orienting of attention', *Quarterly Journal of Experimental Psychology*, 32(1), pp. 3–25.
- Prime, D. J., McDonald, J. J., Green, J. & Ward, L. M. (2008), 'When cross-modal spatial attention fails', *Canadian Journal of Experimental Psychology*, 62(3), pp. 192–197.

- Prokofiev, S. (1936), 'Peter and the Wolf', [Musical Composition].
- Pulkki, V. & Karjalainen, M. (2015), *Communication Acoustics*, Wiley, Chichester, UK.
- Puredata (n.d.), 'Pd-extended', [online], Available:  
<http://puredata.info/downloads/pd-extended>, [Accessed: 2015-02-27].
- Putz, V. (2004), *Spatial Auditory User Interfaces*, Diploma thesis, University of Music and Dramatic Arts Graz, [online], Available: <http://iem.kug.ac.at/en/projects/workspace/projekte-bis-2008/dsp/spatial.html>, [Accessed: 2013-08-23].
- R Core Team (2016), 'R: A language and environment for statistical computing', Vienna, Austria, [online], Available: <https://www.r-project.org/>, [Accessed: 2016-03-30].
- Rai, S. (2015), 'Audio Description App User Trial: Report', Technical report, Royal National Institute of Blind People, London, [online], Available: [http://www.rnib.org.uk/sites/default/files/Audio\\_Description\\_App\\_Trial\\_Report.docx](http://www.rnib.org.uk/sites/default/files/Audio_Description_App_Trial_Report.docx), [Accessed: 2015-09-24].
- Raman, T. V. (1997), *Auditory User Interfaces*, Kluwer Academic Publishers, Boston, MA.
- Rayleigh, L. (1907), 'On our perception of sound direction', *Philosophical Magazine Series* 6, 13(74), pp. 214–232.
- Reisberg, D., Scheiber, R. & Potemken, L. (1981), 'Eye position and the control of auditory attention', *Journal of Experimental Psychology: Human Perception and Performance*, 7(2), pp. 318–323.
- Rhodes, G. (1987), 'Auditory attention and the representation of spatial information', *Perception & Psychophysics*, 42(1), pp. 1–14.
- Roads, C. (1996), *The Computer Music Tutorial*, MIT Press, London, UK.
- Rom, D. M. (1990), 'A sequentially rejective test procedure based on a modified Bonferroni inequality', *Biometrika*, 77(3), pp. 663–665.
- Rorden, C. & Driver, J. (1999), 'Does auditory attention shift in the direction of an upcoming saccade?', *Neuropsychologia*, 37(3), pp. 357–377.
- Rumsey, F. (2001), *Spatial Audio*, Music Technology series, Focal Press, Oxford, UK.
- Ryan, J. (1969), 'Grouping and short-term memory: Different means and patterns of grouping', *Quarterly Journal of Experimental Psychology*, 21(2), pp. 137–147.

- Sachs, M. B. & Kiang, N. Y. S. (1968), 'Two-tone inhibition in auditory-nerve fibers', *The Journal of the Acoustical Society of America*, 43(5), pp. 1120–1128.
- Sætrevik, B. (2012), 'The right ear advantage revisited: speech lateralisation in dichotic listening using consonant-vowel and vowel-consonant syllables', *Laterality: Asymmetries of Body, Brain and Cognition*, 17(1), pp. 119–127.
- Saito, T., Ikei, Y., Hirota, K. & Hirose, M. (2010), 'Spatial voice menu and head gesture interaction system for a wearable computer', in: '20th International Conference on Artificial Reality and Telexistence', pp. 71–76.
- Salamé, P. & Baddeley, A. (1987), 'Noise, unattended speech and short-term memory', *Ergonomics*, 30(8), pp. 1185–1194.
- Salamé, P. & Baddeley, A. (1989), 'Effects of background music on phonological short-term memory', *The Quarterly Journal of Experimental Psychology Section A*, 41(1), pp. 107–122.
- Samsung (n.d.), 'Samsung 88-inch JS9500 Curved Smart 4K 3D SUHD LED TV', [online], Available: <http://www.samsung.com/uk/consumer/tv-audio-video/televisions/suhd-nano-crystal-tvs/UE88JS9500TXXU>, [Accessed: 2016-06-04].
- Scharf, B., Quigley, S., Aoki, C., Peachey, N. & Reeves, A. (1987), 'Focused auditory attention and frequency selectivity', *Perception & Psychophysics*, 42(3), pp. 215–223.
- Schlauch, R. S. & Hafter, E. R. (1991), 'Listening bandwidths and frequency uncertainty in pure-tone signal detection', *The Journal of the Acoustical Society of America*, 90(3), pp. 1332–1339.
- Schmandt, C. (1998), 'Audio hallway: A virtual acoustic environment for browsing', in: 'Proceedings of the 11th Annual ACM Symposium on User Interface Software and Technology', UIST '98, pp. 163–170.
- Schmandt, C. & Mullins, A. (1995), 'AudioStreamer: Exploiting simultaneity for listening', in: 'Proceedings of the SIGCHI Conference on Human Factors in Computing Systems - CHI '95', pp. 218–219.

- Sears, A., Lin, M., Jacko, J. & Xiao, Y. (2003), 'When computers fade... Pervasive computing and situationally-induced impairments and disabilities', in: Jacko, J. & Constantine, S. (Eds.), 'Human Computer Interaction: Theory and Practice', volume 2, pp. 1298–1302.
- Shafiro, V. & Gygi, B. (2007), 'Perceiving the speech of multiple concurrent talkers in a combined divided and selective attention task', *The Journal of the Acoustical Society of America*, 122(6), pp. EL229–235.
- Shakeel, H. & Best, M. (2002), 'Community knowledge sharing: An Internet application to support communications across literacy levels', in: 'IEEE 2002 International Symposium on Technology and Society (ISTAS'02). Social Implications of Information and Communication Technology. Proceedings (Cat. No.02CH37293)', pp. 37–44.
- Shinn-Cunningham, B. G. & Ihlefeld, A. (2004), 'Selective and divided attention: Extracting information from simultaneous sound sources', in: Barrass, S. & Vickers, P. (Eds.), 'Proceedings of the 10th International Conference on Auditory Display (ICAD2004)', [online], Available: <http://www.icad.org/node/2643>, [Accessed: 2013-10-02].
- Shinn-Cunningham, B. G., Schickler, J., Kopčo, N. & Litovsky, R. (2001), 'Spatial unmasking of nearby speech sources in a simulated anechoic environment', *The Journal of the Acoustical Society of America*, 110(2), pp. 1118–1129.
- Shirley, B. & Oldfield, R. (2015), 'Clean audio for TV broadcast: An object-based approach for hearing-impaired viewers', *Journal of the Audio Engineering Society*, 63(4), pp. 245–256.
- Simon-Dack, S. & Teder-Sälejärvi, W. (2008), 'Proprioceptive cues modulate further processing of spatially congruent auditory information. A high-density EEG study', *Brain Research*, 1220, pp. 171–178.
- Simpson, B. D., Brungart, D. S., Iyer, N., Gilkey, R. H. & Hamil, J. T. (2006), 'Detection and localization of speech in the presence of competing speech signals', in: Stockman, T., Nickerson, L. V., Frauenberger, C., Edwards, A. D. N. & Brock, D. (Eds.), 'Proceedings of the 12th International Conference on Auditory Display (ICAD2006)', pp. 129–133.
- Smith, E. E. & Kosslyn, S. M. (2014), *Cognitive Psychology: Mind and Brain*, Pearson Education Limited, Harlow, UK.

- Sodnik, J., Jakus, G. & Tomažič, S. (2011), 'Multiple spatial sounds in hierarchical menu navigation for visually impaired computer users', *International Journal of Human-Computer Studies*, 69(1-2), pp. 100–112.
- Sörqvist, P., Halin, N. & Hygge, S. (2010), 'Individual differences in susceptibility to the effects of speech on reading comprehension', *Applied Cognitive Psychology*, 24(1), pp. 67–76.
- Sparks, D. W. (1976), 'Temporal recognition masking—or interference?', *The Journal of the Acoustical Society of America*, 60(6), pp. 1347–1353.
- Spence, C. & Driver, J. (1994), 'Covert spatial orienting in audition: Exogenous and endogenous mechanisms', *Journal of Experimental Psychology*, 20(3), pp. 555–574.
- Spence, C. & Driver, J. (1996), 'Audiovisual links in endogenous covert spatial attention', *Journal of Experimental Psychology: Human Perception and Performance*, 22(4), pp. 1005–1030.
- Spence, C. & Driver, J. (1997), 'Audiovisual links in exogenous covert spatial orienting', *Perception & Psychophysics*, 59(1), pp. 1–22.
- Spence, C., Ranson, J. & Driver, J. (2000), 'Cross-modal selective attention: On the difficulty of ignoring sounds at the locus of visual attention', *Perception & Psychophysics*, 62(2), pp. 410–424.
- Spence, C. & Read, L. (2003), 'Speech shadowing while driving: On the difficulty of splitting attention between eye and ear', *Psychological Science*, 14(3), pp. 251–256.
- Spieth, W., Curtis, J. F. & Webster, J. C. (1954), 'Responding to one of two simultaneous messages', *The Journal of the Acoustical Society of America*, 26(3), pp. 391–396.
- Stevens, C., Brennan, D. & Parker, S. (2004), 'Simultaneous manipulation of parameters of auditory icons to convey direction, size, and distance: Effects on recognition and interpretation', in: Barrass, S. & Vickers, P. (Eds.), 'Proceedings of the 10th International Conference on Auditory Display (ICAD2004)', [online], Available: <https://smartech.gatech.edu/handle/1853/50916>, [Accessed: 2016-02-18].
- Styles, E. A. (2006), *The Psychology of Attention*, Psychology Press, Hove, UK, 2nd edition.
- Sumby, W. H. & Pollack, I. (1954), 'Visual contribution to speech intelligibility in noise', *The Journal of the Acoustical Society of America*, 26(2), pp. 212–215.

- Summerfield, Q. & Assmann, P. F. (1991), 'Perception of concurrent vowels: Effects of harmonic misalignment and pitch-period asynchrony', *Journal of the Acoustical Society of America*, 89(3), pp. 1364–1377.
- Tajadura-Jiménez, A., Larsson, P., Väljamäe, A., Västfjäll, D. & Kleiner, M. (2010), 'When room size matters: Acoustic influences on emotional responses to sounds', *Emotion*, 10(3), pp. 416–422.
- The Nielson Company (2013), 'Action Figures: How Second Screens are Transforming TV Viewing', [online], Available: <http://www.nielsen.com/us/en/insights/news/2013/action-figures--how-second-screens-are-transforming-tv-viewing.html>, [Accessed: 2014-12-05].
- Theofanos, M. F. & Redish, J. G. (2003), 'Guidelines for accessible - and usable - web sites: Observing users who work with screenreaders', *Interactions*, 10(6), pp. 36–51.
- Thompson, W. F., Schellenberg, E. G. & Letnic, A. K. (2012), 'Fast and loud background music disrupts reading comprehension', *Psychology of Music*, 40(6), pp. 700–708.
- Thurlow, W. R. & Jack, C. E. (1973), 'Certain determinants of the "Ventriloquism Effect"', *Perceptual and Motor Skills*, 36(3), pp. 1171–1184.
- Touchcast (n.d.), 'Touchcast', [online], Available: <http://www.touchcast.com/>, [Accessed: 2013-10-13].
- Treisman, A. M. (1960), 'Contextual cues in selective listening', *Quarterly Journal of Experimental Psychology*, 12(4), pp. 242–248.
- Treisman, A. M. (1964), 'Verbal cues, language, and meaning in selective attention', *The American Journal of Psychology*, 77(2), pp. 206–219.
- Tun, P. A., O'Kane, G. & Wingfield, A. (2002), 'Distraction by competing speech in young and older adult listeners', *Psychology and Aging*, 17(3), pp. 453–467.
- U.S. Government (2013a), 'User interfaces provided by digital apparatus', in: 'Title 47 Code of Federal Regulations', p. \$79.107, U.S. Government Publishing Office.
- U.S. Government (2013b), 'Video programming guides and menus provided by navigation devices', in: 'Title 47 Code of Federal Regulations', p. \$79.108, U.S. Government Publishing Office.

- USwitch (n.d.), 'IPTV via Broadband - uSwitch Broadband's Guide to IPTV', 'uSwitch.com', [online], Available: [http://www.uswitch.com/broadband/guides/what\\_is\\_ip\\_tv/](http://www.uswitch.com/broadband/guides/what_is_ip_tv/), [Accessed: 2016-06-01].
- van Alphen, P. & McQueen, J. M. (2001), 'The time-limited influence of sentential context on function word identification', *Journal of Experimental Psychology: Human Perception and Performance*, 27(5), pp. 1057–1071.
- van Beem, A. (2013), 'Old Philips television set, pic6.JPG', 'Wikimedia Commons', [online], Available: [https://commons.wikimedia.org/wiki/File:Old\\_Philips\\_television\\_set,\\_pic6.JPG](https://commons.wikimedia.org/wiki/File:Old_Philips_television_set,_pic6.JPG), [Accessed: 2016-08-02].
- Vargas, M. L. M. & Anderson, S. (2003), 'Combining speech and earcons to assist menu navigation', in: Brazil, E. & Shinn-Cunningham, B. G. (Eds.), 'Proceedings of the 2003 International Conference on Auditory Display', pp. 38–41.
- Vazquez-Alvarez, Y. & Brewster, S. A. (2010), 'Designing spatial audio interfaces to support multiple audio streams', in: 'Proceedings of the 12th International Conference on Human Computer Interaction with Mobile Devices and Services - MobileHCI '10', pp. 253–256.
- Vazquez-Alvarez, Y. & Brewster, S. A. (2011), 'Eyes-free multitasking: The effect of cognitive load on mobile spatial interfaces', in: 'Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems - CHI '11', pp. 2173–2176.
- Vinayagamorthy, V., Allen, P., Hammond, M. & Evans, M. (2012), 'Researching the user experience for connected TV: A case study', in: 'CHI '12 Extended Abstracts on Human Factors in Computing Systems', CHI EA '12, pp. 589–604.
- W3C (2014), 'HTML5: A Vocabulary and Associated APIs for HTML and XHTML', [online], Available: <https://www.w3.org/TR/html5/single-page.html>, [Accessed: 2015-07-14].
- Walczak, A. (2016), 'Foreign language class with audio description: A case study', in: Matamala, A. & Orero, P. (Eds.), 'Researching Audio Description: New Approaches', Palgrave Studies in Translating and Interpreting, chapter 10, pp. 187–204, Palgrave Macmillan UK, London.
- Walker, A., Brewster, S., McGookin, D. & Ng, A. (2001), 'Diary in the sky: A spatial audio display for a mobile calendar', in: Blandford, A., Vanderdonckt, J. & Gray, P. (Eds.),

- 'People and Computers XV—Interaction Without Frontiers', pp. 531–539, Springer, London.
- Walker, B. N., Lindsay, J., Nance, A., Nakano, Y., Palladino, D. K., Dingler, T. & Jeon, M. (2013), 'Spearcons (speech-based earcons) improve navigation performance in advanced auditory menus', *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 55(1), pp. 157–182.
- Walker, B. N., Nance, A. & Lindsay, J. (2006), 'Spearcons: Speech-based earcons improve navigation performance in auditory menus', in: 'Proceedings of the 12th International Conference on Auditory Display (ICAD2006)', pp. 63–68.
- Wallach, H. (1940), 'The role of head movements and vestibular and visual cues in sound localization', *Journal of Experimental Psychology*, 27(4), pp. 339–368.
- Ward, L. M. (1994), 'Supramodal and modality-specific mechanisms for stimulus-driven shifts of auditory and visual attention', *Canadian Journal of Experimental Psychology*, 48(2), pp. 242–259.
- Warren, R. M. (1970), 'Perceptual restoration of missing speech sounds', *Science*, 167(3917), pp. 392–393.
- Watson, C. (2006), 'Untitled contribution', in: 'Computational and Systems Neuroscience Workshop: Difficult Issues in Auditory Scene Analysis', [online], Available: [http://www.isr.umd.edu/Labs/NSL/Cosyne/masking\\_counterpoint.htm](http://www.isr.umd.edu/Labs/NSL/Cosyne/masking_counterpoint.htm), [Accessed: 2013-09-09].
- Webster, J. C. & Thompson, P. O. (1954), 'Responding to both of two overlapping messages', *The Journal of the Acoustical Society of America*, 26(3), pp. 396–402.
- Werner, S., Hauck, C., Roome, N., Hoover, C. & Choates, D. (2015), 'Can VoiceScapes assist in menu navigation?', in: 'Proceedings of the Human Factors and Ergonomics Society 59th Annual Meeting', 2010, pp. 1095–1099.
- Wersényi, G. (2009), 'Evaluation of auditory representations for selected applications of a graphical user interface', in: 'Proceedings of the 15th International Conference on Auditory Display (ICAD2009)', [online], Available: <http://hdl.handle.net/1853/51294>, [Accessed: 2012-10-05].



- Wexler, B. E. & Halwes, T. (1983), 'Increasing the power of dichotic methods: The fused rhymed words test', *Neuropsychologia*, 21(1), pp. 59–66.
- Wickens, C. D. (1980), 'The structure of attentional resources', in: Nickerson, R. S. (Ed.), 'Attention and Performance VIII', chapter 12, pp. 239–258, Lawrence Erlbaum Associates, Hillsdale.
- Wickens, C. D. (1984), 'Processing resources in attention', in: Parasuraman, R. & Davies, D. R. (Eds.), 'Varieties of Attention', pp. 63–102, Academic Press, New York, NY.
- Wickens, C. D. (2002), 'Multiple resources and performance prediction', *Theoretical Issues in Ergonomics Science*, 3(2), pp. 159–177.
- Wickens, C. D. (2008), 'Multiple resources and mental workload', *Human Factors: The Journal of the Human Factors and Ergonomics*, 50(3), pp. 449–455.
- Wickens, C. D., Hollands, J. G., Banbury, S. & Parasuraman, R. (2016), *Engineering Psychology and Human Performance*, Routledge, Oxon, UK, 4th edition.
- Wikipedia Contributors (2017), 'Gray Langur', 'Wikipedia', [online], Available: [https://en.wikipedia.org/wiki/Gray\\_langur](https://en.wikipedia.org/wiki/Gray_langur), [Accessed: 2017-01-22].
- Wilcox, R. R. (2012), *Introduction to Robust Estimation & Hypothesis Testing*, Academic Press, San Diego, CA, 3rd edition.
- Wilcox, R. R. (2016), 'Rallfun-v31', [online], Available: <http://dornsife.usc.edu/assets/sites/239/docs/Rallfun-v31.txt>, [Accessed: 2016-09-09].
- Williams, T. (2013), 'New Antiques Roadshow Play-along App', 'BBC Internet Blog', [online], Available: [http://www.bbc.co.uk/blogs/internet/posts/antiques\\_roadshow\\_play-along\\_a](http://www.bbc.co.uk/blogs/internet/posts/antiques_roadshow_play-along_a), [Accessed: 2013-10-17].
- Witkin, H. A., Wapner, S. & Leventhal, T. (1952), 'Sound localization with conflicting visual and auditory cues', *Journal of Experimental Psychology*, 43(1), pp. 58–67.
- Wohn, D. Y. & Na, E.-K. (2011), 'Tweeting about TV: Sharing television viewing experiences via social media message streams', *First Monday*, 16(3), [online], Available: <http://firstmonday.org/ojs/index.php/fm/article/view/3368/2779>, [Accessed: 2015-10-19].

- Wood, N. & Cowen, N. (1995), 'The cocktail party phenomenon revisited: How frequent are attention shifts to one's name in an irrelevant auditory channel?', *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 21(1), pp. 255–260.
- Woods, R. L. & Satgunam, P. (2011), 'Television, computer and portable display device use by people with central vision impairment', *Ophthalmic & Physiological Optics: The Journal of the British College of Ophthalmic Opticians (Optometrists)*, 31(3), pp. 258–274.
- World Health Organization (2014), 'Visual impairment and blindness', World Health Organization, [online], Available: <http://www.who.int/mediacentre/factsheets/fs282/en/>, [Accessed: 2015-10-30].
- Worley, J. W. & Darwin, C. J. (2002), 'Auditory attention based on differences in median vertical plane position', in: 'Proceedings of the 2002 International Conference on Auditory Display', [online], Available: <https://smartech.gatech.edu/handle/1853/51344>, [Accessed: 2016-07-25].
- Yalla, P. & Walker, B. N. (2007), 'Advanced Auditory Menus', Technical report, Georgia Institute of Technology Gvu Center, [online], Available: <http://sonify.psych.gatech.edu/publications/pdfs/2007-GVU0712-YallaWalker.pdf>, [Accessed: 2012-10-24].
- Yalla, P. & Walker, B. N. (2008), 'Advanced auditory menus: Design and evaluation of auditory scroll bars', in: 'Proceedings of the Annual ACM Conference on Assistive Technologies (ASSETS'08)', pp. 105–112.
- Ylias, G. & Heaven, P. C. (2003), 'The influence of distraction on reading comprehension: A big five analysis', *Personality and Individual Differences*, 34(6), pp. 1069–1079.
- Yost, W. A. (2006), 'Informational masking: What is it?', in: 'Computational and Systems Neuroscience Workshop: Difficult issues in auditory scene analysis', [online], Available: <http://www.isr.umd.edu/Labs/NSL/Cosyne/Yost.htm>, [Accessed: 2013-02-12].
- YouView (n.d.), 'YouView Features', [online], Available: <http://www.youview.com/features/>, [Accessed: 2015-10-18].
- Zeger, S. L. & Liang, K.-Y. (1986), 'Longitudinal data analysis for discrete and continuous outcomes', *Biometrics*, 42(1), pp. 121–130.

- Zwicker, U. T. (1984), 'Auditory recognition of diotic and dichotic vowel pairs', *Speech Communication*, 3(4), pp. 265–277.