Perceptual Plasticity in Adverse Listening Conditions: Factors Affecting Adaptation to Accented and Noise-Vocoded Speech

A thesis submitted to the University of Manchester

for the degree of Doctor of Philosophy

in the Faculty of Medical and Human Sciences

2015

Briony Banks

School of Psychological Sciences

Table of Contents

List of Tables

Chapter 3

Chapter 4

Chapter 5

List of Figures

Chapter 3

Chapter 4

Glossary of Key Terms and Abbreviations

Fixation	A period of relatively stable eye gaze in between saccades (eye movements).
IA	Interest Area (used in the spatial analysis of eye movements)
IEEE	Institute of Electrical and Electronics Engineers
Lexical	Relating to the words or vocabulary of a language.
Lexical item	A single word.
Noise-vocoded speech	Spectrally-degraded speech.
Non-words	Phonologically correct but semantically meaningless words, e.g. *fusher* or *breeg*.
Phoneme	A perceptually distinct unit of sound in a specified language that distinguish one word from another, for example *p* and *b* in the English words *pad* and *bad*.
Phonetic	Relating to speech sounds.
Phonological	Relating to the system of contrastive relationships among speech sounds that constitute the fundamental components of a language.
Place of articulation	When articulating a consonant: the point of contact between two articulators, for example the lips closing or the tongue touching the teeth.
Saccade	Rapid eye movement from one spatial location to another.
Semantic	Relating to the meaning in language.
SNR	Signal-to-noise ratio.
Speech signal	The acoustic patterns of speech.
SRT	Speech Reception Threshold.
Syntactic	Relating to the arrangement of words and phrases to create grammatical sentences in a language.

Abstract

The University of Manchester
Briony Banks
Doctor of Philosophy

Perceptual Plasticity in Adverse Listening Conditions: Factors Affecting Adaptation to Accented and Noise-vocoded Speech

September 2015

Adverse listening conditions can be a hindrance to communication, but humans are remarkably adept at overcoming them. Research has begun to uncover the cognitive and behavioural mechanisms behind this perceptual plasticity, but we still do not fully understand the reasons for variability in individual responses. The research reported in this thesis addressed several factors which would further this understanding.

Study 1 examined the role of cognitive ability in recognition of, and perceptual adaptation to, accented speech. A measure of executive function predicted greater and more rapid perceptual adaptation. Vocabulary knowledge predicted overall recognition of the accented speech, and mediated the relationship between working memory and recognition accuracy. Study 2 compared recognition of, and perceptual adaptation to, accented speech with and without audiovisual cues. The presence of audiovisual cues improved recognition of the accented speech in noise, but not perceptual adaptation. Study 3 investigated when perceivers make use of visual speech cues during recognition of, and perceptual adaptation to, audiovisual noise-vocoded speech. Listeners' eye gaze was analysed over time and related to their performance. The percentage and duration of fixations on the speaker's mouth increased during recognition of individual sentences, while the duration of fixations on the mouth decreased as perceivers adapted to the noise-vocoded speech over the course of the experiment. Longer fixations on the speaker's mouth were related to better speech recognition.

Results demonstrate that perceptual plasticity of unfamiliar speech is driven by cognitive processes, but can also be modified by the modality of speech (audiovisual or audio-only). Behavioural responses, such as eye gaze, are also related to our ability to respond to adverse conditions. Speech recognition and perceptual adaptation were differentially related to the factors in each study and therefore likely reflect different processes; these measures should therefore both be considered in studies investigating listeners' response to adverse conditions. Overall, the research adds to our understanding of the mechanisms and behaviours involved in perceptual plasticity in adverse listening conditions.

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification, of this or any other university, or other institute of learning.

Copyright Statement

Acknowledgements

A big, big thank you to Patti for the opportunity, and all her support and encouragement over the past four years, especially for still being there for me after moving down south.

A big thank you also to Emma and Kevin for all their kind support and help.

Thank you to the BEAM lab for support, friendship and lots of helpful journal club debates!

Thank you to Ellen for helpful advice and some very enjoyable experience in public engagement.

Thanks to Taiji and Hannah for help recording the stimuli.

Thank you to all my friends and family for encouragement, positive words, much-needed tea breaks (Jen and Christina!), and for keeping me (almost) sane.

An especially big thank you to Chris, for persuading me that this was actually a good idea; for having faith in me, providing endless reassurance and helping me believe in myself; for living without me for 3 years and travelling 200 miles every month to see me; and, particularly, for not killing me after spending three days shut in a windowless room trying to speak my novel accent.

CHAPTER 1.
INTRODUCTION AND LITERATURE REVIEW

1.1 Introduction

Adverse listening conditions can be defined as any context that reduces the success or ease of speech recognition. Different categories of adverse condition have been proposed in the literature, relating them to the speech source, the environment or to the receiver (Mattys, Davis, Bradlow, & Scott, 2012), for example, an unfamiliar accent, background noise or a hearing impairment. All types of adverse condition can be frequently encountered in everyday life and pose a challenge to successful communication, particularly in modern society where communication often takes place in less-than-ideal conditions. Research within this area thus has implications for both healthy and clinical populations. Perhaps for this reason, there is currently great interest among researchers in how listeners adapt to, and compensate for, adverse listening conditions. This interdisciplinary field has gained increasing interest from psychologists, linguists, audiologists and neuroscientists, who wish to increase our understanding of how we perceive and process speech, and how we adapt and respond to challenges during the process. Current and past research has addressed many different types of adverse condition from all three categories; it has investigated how different listening conditions affect speech recognition, as well as the strategies listeners use to compensate for them, and the sensory, cognitive and neural processes involved in responding and adapting to them. Previously, models of speech processing focused solely on successful speech communication in ideal conditions, but researchers are increasingly aware that dealing with adverse conditions is an integral part of human communication, and are now acknowledging their place within these models. However, much is still to be learnt, particularly at the cognitive and neural levels, as well as identifying ways in which the negative effects of adverse conditions can be lessened or overcome.

Humans deal with adverse listening conditions primarily through behavioural (and consequently neural) adaptation – that is, through perceptual plasticity. Plasticity in

the wider sense refers to changes at the behavioural and neural level of an organism in response to changes in the environment (Goldstone, 1998); in the context of this thesis, plasticity corresponds to our response to adverse listening conditions. Although such conditions make communication more difficult or effortful, we are often able to adapt to, or compensate for them in order to successfully recognise and understand speech; this might be compensating for background noise by looking at a speaker's face, or 'tuning in' to the patterns of an unfamiliar accent (see Cristia et al., 2012; Samuel & Kraljic, 2009 for reviews). Although perceptual plasticity of speech has been quite widely studied, we still do not fully understand how it occurs, or whether particular strategies and behaviours can improve it. Studying the common factors that affect the success of perceptual plasticity may help us to understand the mechanisms behind it, as well as how to potentially improve communication in adverse listening conditions.

This thesis presents a series of experiments which investigated perceptual plasticity of speech under adverse listening conditions. Specifically, the work addressed two factors that affect speech recognition when the source of the speech signal itself is unclear: i) cognitive ability, and ii) the availability and use of audiovisual speech cues. The work has assessed the extent to which these factors affect perceptual plasticity of unfamiliar speech, with the overall aim of better understanding the underlying processes of perceptual plasticity, and identifying whether different strategies can improve listeners' response to adverse listening conditions.

The thesis is written in the alternative format, meaning that each experimental chapter is presented in the style of a journal manuscript. This format is particularly suitable for the work undertaken here, as each chapter comprises an experiment or a set of experiments addressing a different factor relating to the overall topic. As a consequence, some details of the experimental work such as the justification, methods and implications may be duplicated within the thesis. Details of each subsequent chapter are presented in the following paragraphs, including publications and conference presentations resulting from the work.

1.1.2 Outline of chapters included in the thesis

Chapter 2. General Methods. This chapter presents the main methodological considerations for the experimental work carried out, including participants, stimuli, and tests conducted. Specific details relating to each study are also provided in the relevant experimental chapters.

Chapter 3. Cognitive predictors of perceptual adaptation to accented speech. This study addressed the role of cognitive ability in recognition of, and adaptation to, accented speech. Specifically, it assessed the contribution of executive function, vocabulary knowledge and working memory. The manuscript was accepted for publication in the Journal of the Acoustical Society of America in March 2015.

Chapter 4. Audiovisual cues benefit recognition of accented speech in noise but not perceptual adaptation. This chapter comprises two studies (hereafter referred to as Study 1 and Study 2) investigating whether the availability of audiovisual speech cues benefitted perceptual adaptation to accented speech in noise. Specifically, each study tested whether perceptual adaptation to accented speech was greater with the presence of audiovisual speech cues (that is, when the listener is face-to-face with the speaker) compared to audio-only speech cues. The manuscript was accepted for publication in Frontiers of Human Neuroscience for a special topic on accented speech in July 2015.

Chapter 5. Eye gaze during recognition of audiovisual noise-vocoded speech. The final study investigated when listeners use visual speech cues during recognition of audiovisual noise-vocoded sentences, and during perceptual adaptation to accented speech, by analysing eye gaze towards a speaker's eyes and mouth during recognition of i) individual sentences, and ii) multiple sentences (over the course of the experiment). As a secondary aim, the relationship between eye gaze and recognition of audiovisual noise-vocoded speech was also investigated. At the time of writing, this manuscript was being prepared for submission to the Journal of Experimental Psychology: Human Perception and Performance.

Chapter 6. General Discussion and Conclusion. The final chapter discusses the novelty and implications of the experimental work within the wider research

context, as well as discussing the main limitations of the work. Directions for related future research are also outlined.

1.1.3 Journal articles resulting from the thesis

Banks, B., Gowen, E., Munro, K.J., Adank, P. (2015). Cognitive predictors of perceptual adaptation to accented speech. Journal of the Acoustical Society of America, 137 (4), 2015-2024.

Banks, B., Gowen, E., Munro, K.J., Adank, P. (2015). Audiovisual cues benefit recognition of accented speech in noise but not perceptual adaptation. Frontiers in Human Neuroscience. 9:422. doi: 10.3389/fnhum.2015.00422

Banks, B., Gowen, E., Munro, K.J., Adank, P. (In prep.) Eye gaze during recognition of audiovisual noise-vocoded speech. Journal of Experimental Psychology: Human Perception and Performance.

1.1.4 Oral presentations resulting from the thesis

Eye gaze during comprehension of audiovisual speech in adverse listening conditions. UCL Speech Science Forum, 2014, London, UK

Eye gaze during comprehension of audiovisual speech in adverse listening conditions. Experimental Psychology Society Spring Meeting, 2014, Kent, UK.

1.1.5 Poster presentations resulting from the thesis

Can offline audiovisual training aid perceptual adaptation to accented speech? Neurobiology of Language, 2012, San Sebastian, Spain

Cognitive predictors of perceptual adaptation to accented speech in noise. Cognitive Hearing Science, 2013, Linkoping, Sweden

Cognitive predictors of perceptual adaptation to accented speech in noise. British Society of Audiology, 2013, Keele, UK

Eye movements during perception of audiovisual unfamiliar speech. British Oculomotor Group, 2013, Manchester, UK

Cognitive Predictors of Perceptual Adaptation to Accented Speech in Noise. Neuroscience Research Institute, 2013, Manchester, UK

Eye gaze during perceptual adaptation of audiovisual speech in adverse listening conditions. Neurobiology of Language, 2014, Amsterdam, Netherlands

Individual differences in eye gaze during audiovisual sentence recognition. Cognitive Hearing Science, Linkoping, Sweden

Dr Patti Adank, Dr Emma Gowen and Professor Kevin Munro have co-authored all publications. However, the design of all experiments and stimuli, data collection and analysis was conducted by Briony Banks, as was the writing of all material presented in the thesis.

1.2 Literature Review

The following literature review introduces the general themes of perceptual plasticity of speech and adverse listening conditions, as they have been addressed in the literature. It particularly discusses research into adverse conditions which relate to the speech source, as these particular conditions are the subject of this thesis. The review then focuses primarily on perceptual plasticity in response to source-related adverse conditions, discussing current knowledge of the nature, outcomes and driving mechanisms of perceptual plasticity. Finally, there is a discussion of the variety of factors that affect perceptual plasticity, with a particular focus on the two factors to be investigated in this thesis: cognitive ability and audiovisual speech cues.

1.2.1 Perceptual plasticity of speech

Perceptual plasticity of speech (that is, how listeners respond to adverse listening conditions) demonstrates the remarkable flexibility of the human perceptual system; we are able to understand different speakers when listening conditions are challenging, even if at first the speech seems incomprehensible. Research has repeatedly

shown that perceptual plasticity is a robust and lifelong ability (e.g. see Cristia et al., 2012 for a review), present from as early as 19 months of age (Schmale, Cristia, & Seidl, 2012; White & Aslin, 2011), and continuing into old age (Adank & Janse, 2010; Peelle & Wingfield, 2005). Perceptual plasticity has been investigated in terms of our overt behavioural responses, such as looking at a speaker's facial movements when listening in a noisy environment, or in terms of the perceptual, cognitive and neural processes behind it. Perceptual plasticity can be studied using variations that naturally occur in speech, for example a non-native accent, or it can be studied using artificial manipulations, such as spectrally-altered or time-compressed speech. In an experimental setting, perceptual plasticity is commonly investigated by measuring the success of speech communication, and this can be defined in two ways: 1) the immediate or overall recognition of speech, for example words, sentences or longer passages; or 2) the amount or rate of improvement in speech recognition over a period of time. Within the literature, the first definition is normally termed 'speech recognition' (e.g. Shannon, Zeng, Kamath, Wygonski, & Ekelid, 1995) while the second is referred to (often interchangeably) as 'perceptual adaptation' (e.g. Janse & Adank, 2012) or 'perceptual learning' (e.g. Peelle & Wingfield, 2005).

These definitions have largely been studied independently of one another, but with the underlying assumption that they both represent types of perceptual plasticity. As both definitions reflect how listeners deal with adverse listening conditions, they will both be discussed within this thesis. Specifically, 'speech recognition' will denote overall speech recognition or speech recognition at a given point in time, whereas 'perceptual adaptation' will denote any improvement in speech recognition over a given period of time. The term 'perceptual plasticity' will be used as a more general term to indicate any response to adverse listening conditions.

1.2.2 Adverse Listening Conditions

A variety of adverse listening conditions have been studied in relation to perceptual plasticity, but to different extents. Out of the three definitions described by Mattys et al (2012), conditions relating to the environment (such as the presence of background noise) have been extensively studied over the past few decades (for example, see Akeroyd, 2008), perhaps because they are the most well-known and

commonly encountered of all adverse conditions, and because they are particularly problematic for users of hearing aids and individuals with hearing loss. For example, research has investigated the effects of different types and levels of noise on speech recognition (e.g. Miller, 1947; Rosen, Souza, Ekelund, & Majeed, 2013); how listeners compensate for and overcome the presence of background noise (e.g. Pichora-Fuller, 2003; Sumby & Pollack, 1954); and, more recently, the 'effort' involved in processing speech in this context (Zekveld, Kramer, & Festen, 2010). Consequently, a great deal is known about the effects of environmental adverse conditions. Similarly, there has been much research, particularly from a clinical perspective, into the effects of hearing loss on speech recognition (e.g. Moore, 1998; Tyler, Summerfield, Wood, & Fernandes, 1982). Whilst it is unarguably important for such research to take place, for example to improve hearing devices, source-related adverse listening conditions have received less attention in comparison.

Variation in the speech signal, such as an unfamiliar accent, was historically viewed by researchers as 'noise' that listeners had to normalise in order to successfully recognise and understand the underlying speech (Pisoni, 1997). Variation in the speech source was therefore not widely studied as an adverse listening condition in its own right. This view has more recently been challenged, for example by exemplar theories of speech recognition (Goldinger, 1996; Hawkins, 2003). Consequently, the study of variation in the speech source has become a topic of interest in recent years, and research into source-related adverse listening conditions has developed considerably. Understanding such conditions is pertinent to our daily lives, for example encountering speech with an unfamiliar accent is increasingly common. Furthermore, source-related conditions can interact with other types of adverse condition; for example, recognising accented speech poses problems for certain populations, such as older listeners with hearing impairments (Adank & Janse, 2010), or individuals with cognitive impairments (Hailstone et al., 2012). Studying adverse conditions relating to the speech source can provide valuable insight into how listeners, from healthy or clinical populations, adapt to common variation in the speech signal and to variable listening conditions, and is essential for forming comprehensive models of speech processing.

1.2.3 Source-related adverse listening conditions

Perhaps the most commonly encountered adverse listening condition related to the speech source, is accented speech (Adank & Janse, 2010; Bradlow & Bent, 2008; Clarke & Garrett, 2004). This might be non-native (e.g. Chinese), regional (e.g. Scottish or Northern British English), or, in an experimental setting, a novel accent (Adank & Janse, 2010; Janse & Adank, 2012; Maye, Aslin, & Tanenhaus, 2008). Regardless of the type of accent, the common denominator is that the listener is unfamiliar with the patterns of the particular variant. Accented speech comprises variation in the phonetic and prosodic patterns of speech, which differ from the listener's 'normal' representations or expectations; for example, a Northern British English speaker who consistently produces the phoneme /a/ within the word 'grass', might perceive Southern British English, which uses the phoneme /ɑː/ within the same word, as an unfamiliar accent. A regional accent comprises phonetic variation different to the standard form, according to geographic identity, but this term usually refers to an accent from a region where the same language is spoken (e.g. Scottish). A speaker with a non-native accent, however, can introduce further variation into the speech signal by introducing the phonetic patterns of their own native language; for example, a Japanese speaker may pronounce the English phonemes /l/ or /r/ as the Japanese intermediate phoneme /ɽ/, which is likely unfamiliar to a native British English speaker. A novel accent can potentially be constructed according to either of these patterns, by systematically varying the phonetic patterns of a language by substituting particular phonemes for native or non-native variants.

Aside from accented speech, other natural speech variations that have been addressed by the literature are dysarthric speech – that is, the unclear, variable speech produced by individuals with dysarthria, a neurological disorder that impairs the speech motor system (Borrie, McAuliffe, & Liss, 2012), and speech produced by deaf speakers (Boothroyd, 2010; Mcgarr, 1983). Researchers have also extensively used artificial acoustic distortions not encountered in natural listening contexts. This has included time-compressed speech (e.g. Dupoux & Green, 1997), whereby speech is artificially compressed to present the listener with a 'speeded-up' speech stimulus; synthetic speech (Schwab, Nusbaum, & Pisoni, 1985), and spectrally-rotated speech, which inverts spectral characteristics of speech (Green, Rosen, Faulkner, & Paterson, 2013).

These manipulations provide a stimulus that is clearly recognisable as speech, but is not necessarily intelligible on first listen.

The most commonly used acoustic manipulation is noise-vocoded speech (Davis, Johnsrude, Hervais-Adelman, Taylor, & McGettigan, 2005; Faulkner, Rosen, & Smith, 2000; Hervais-Adelman, Davis, Johnsrude, & Carlyon, 2008; Shannon et al., 1995). This distortion alters the spectral information in the speech signal, resulting in speech that sounds like a 'noisy whisper', and forces the listener to rely more on temporal cues or changes in amplitude cues (that is, loudness and softness). An advantage of such acoustic manipulations is that they can be created at varying degrees of intelligibility depending on the requirements of the experiment, for example noise-vocoded speech is made more intelligible by using a greater number of frequency bands to create it (Davis & Johnsrude, 2003; Loizou, Mani, & Dorman, 2003; Shannon et al., 1995).

The type of adverse condition used in any experiment depends on the particular aims of the study. Investigating dysarthric speech may be important in a clinical setting to potentially improve rehabilitation or communication with patients, while accented speech can help us to understand how we learn to adapt to unfamiliar perceptual patterns, as well as improving communication in clinical settings where patients and clinicians have differing accents, for example. Spectrally-rotated speech has been used in neuroimaging studies to differentiate between brain regions activated by intelligible and non-intelligible speech (Scott, Blank, Rosen, & Wise, 2000), whereas noise-vocoded speech was created to simulate hearing with a cochlear implant, and has thus been the subject of studies investigating hearing with these devices (Faulkner et al., 2000; Shannon et al., 1995). All of the conditions discussed here have a similarly negative effect on speech recognition, and therefore provide contexts from which to study perceptual plasticity. However, acoustic and other artificial distortions offer greater control over the degree of intelligibility and type of variation, and therefore can provide tighter experimental control. Nevertheless, they are not naturally encountered in everyday life. For research of purely scientific or theoretical interest, it could be argued that more ecologically valid conditions such as accented speech are more important than artificial distortions, as these are the conditions that are encountered frequently outside of the laboratory and thus reflect natural processes of perceptual plasticity. Moreover,

we do not know if the same cognitive and perceptual mechanisms drive adaptation to natural and artificial variations, as no such comparison has been made to date. The speech variation used in any particular study should therefore be considered carefully, and responses to a range of speech variations need to be investigated using similar methods, in order to compare differences between them.

1.2.4 Perceptual plasticity of speech under source-related adverse listening conditions

Encountering an unfamiliar speech source generally affects both the success and ease of recognition. Listening to an unfamiliar regional or novel accent, for example, results in poorer speech recognition compared to familiar-accented speech (Adank, Evans, Stuart-Smith, & Scott, 2009; Adank & Janse, 2010). Regional or non-native accented speech is also processed more slowly than familiar, native-accented speech (Clarke & Garrett, 2004; Floccia, Goslin, Girard, & Konopczynski, 2006), while processing noise-vocoded speech is more effortful than clear speech (Wild et al., 2012). The relative effects of a source-related adverse condition depend on many factors, such as the baseline level of intelligibility (e.g. Bradlow & Bent, 2008; see section 1.2.7.1 for a full discussion). Nevertheless, listeners can invariably adapt to the different types of unfamiliar speech – that is, their speech recognition will improve over time due to increased exposure.

Single experiments have demonstrated that recognition accuracy can quickly improve by around 10% for accented speech (Bradlow & Bent, 2008), 10-15% for time-compressed speech (Dupoux & Green, 1997), and around 30% for noise-vocoded speech (Davis et al., 2005). Perceptual adaptation tends to be rapid, with listeners' performance improving significantly with exposure to around twenty sentences; this has been observed for time-compressed (Dupoux & Green, 1997; Pallier, Sebastian-Galles, Dupoux, Christophe, & Mehler, 1998), foreign-accented (Clarke & Garrett, 2004; Gordon-Salant, Yeni-Komshian, & Fitzgibbons, 2010), and noise-vocoded speech (Davis et al., 2005). This rapid adaptation perhaps reflects the common sensation of 'tuning in' when listening to an unfamiliar speaker, whereby an initially unintelligible accent can quickly become intelligible to the listener. Adaptation is a robust ability; for example, interruptions in stimuli presentation or changes in the compression rate of

time-compressed speech do not affect the overall amount of adaptation (Dupoux & Green, 1997; Golomb, Peelle, & Wingfield, 2007). Adaptation can also occur after exposure to single phonemes (Hazan, Sennema, Iba, & Faulkner, 2005), words (Greenspan, Nusbaum, & Pisoni, 1988; Hervais-Adelman et al., 2008), or sentences (e.g. Clarke & Garrett, 2004; Dupoux & Green, 1997), regardless of the baseline intelligibility of the speaker (Bradlow & Bent, 2008; Davis et al., 2005) and, with adaptation to accented speech, regardless of the type of accent (Pinet, Iverson, & Evans, 2011).

Most studies of perceptual adaptation have focused on the short-term effects of exposure to unfamiliar speech (that is, effects observed during a single testing session), or on improvements with trained items; however, some long-term and generalised adaptation effects have also been observed. For example, adaptation that has occurred following training with a closed-set of noise-vocoded words (that is, training with specific words only), can generalise to new, untrained words (Davis & Johnsrude, 2003); similarly, training with multiple accents can lead to better recognition of untrained accents (Baese-Berk, Bradlow, & Wright, 2013). Long term retention of learning has been shown with perceptual adaptation to non-native phonemes; for example, native Japanese speakers were able to accurately distinguish the British English /r/ /l/ phonemic contrast three months after initial perceptual training (Lively, Pisoni, Yamada, Tohkura, & Yamada, 1994). Similar training has even led to improvements in production of the same phonemic contrast (Bradlow, Akahane-Yamada, Pisoni, & Tohkura, 1999; Bradlow, Pisoni, AkahaneYamada, & Tohkura, 1997). Nevertheless, the majority of perceptual adaptation studies have not examined the retention or generalisation of learning effects past one or two experimental sessions. This is perhaps a limitation of the current field of research, as the long-term effects of perceptual adaptation observed in the laboratory are not generally known. To listeners, long-term retention of learning may be a key aspect of successful communication, and would be particularly pertinent to individuals who have difficulty adapting to adverse conditions in the short-term such as older adults (Adank & Janse, 2010).

1.2.5 The outcomes of perceptual adaptation

Changes to the perceptual system that take place during and following perceptual adaptation have been of great interest to researchers over the last two decades; that is, how an unfamiliar speech variant is encoded in memory. There is a growing body of evidence that adapting to accented or unclear speech essentially changes how we perceive and categorise phonemes (Eisner & McQueen, 2005, 2006; Kraljic & Samuel, 2006, 2011; McQueen, Norris, & Cutler, 2006; Norris, McQueen, & Cutler, 2003). For example, when listeners are exposed to an ambiguous sound [?] somewhere between [f] and [s], in the context of words ending in [f] (e.g. "wol?"), they are more likely to subsequently categorise the sound as [f] when the sound is presented in isolation (Norris et al., 2003). Such studies provide evidence that perceptually adapting to an unfamiliar accent changes our perceptual categorisation of phonemes, by 'shifting' the boundaries of each phonemic category to include the new speech sound. This neatly explains what happens when we adapt to the phonetic variation in accented or unclear (for example, dysarthric) speech. However, it does not explain how we encode and recognise an entire accent, which comprises a consistent pattern of phonemic variation that one can ultimately learn and recognise as, for example, a Scottish or French accent. This process is likely to involve pattern learning as well as shifts in perceptual boundaries, but this has not yet been investigated in the literature – that is, the distinction between encoding speech variation at the phonemic level and at the whole accent level, has not yet been clarified.

It is also unclear how the theory extends to other types of unfamiliar speech, such as adaptation to acoustic distortions such as noise-vocoded speech. These variants comprise an acoustic degradation of the entire speech signal rather than variation at the phonemic level, and adapting to them may therefore result in different perceptual changes in the listener. A possible theory of perceptual adaptation to such distortions is that a listener's attention is retuned. In this way, the listener 'fine-tunes' their attention to the acoustic properties that are left intact and perceptible in the degraded signal (for example, the temporal cues in noise-vocoded speech), but that are normally not relied upon for speech recognition when the speech is clear and unaltered. This fits in with a theory of 're-tuned attention' proposed by Amitay (2009), based on the reverse Hierarchy Theory of visual perceptual learning (Ahissar & Hochstein, 2004), and

supported by evidence of lower-level auditory training that is also modified by attention (Halliday, Moore, Taylor, & Amitay, 2011). However, we do not know exactly how attention is retuned – for example, how do listeners learn which acoustic properties to attune their attention to, and how does this affect internal representations of speech? The nature of adaptation to such acoustic distortions is therefore still not fully understood.

1.2.6 The driving mechanisms of adaptation

Mounting evidence suggests that recognition of unfamiliar speech, and perceptual adaptation, are driven primarily by high-level linguistic and cognitive processes, with listeners relying on external information from the environment and speech context, as well as internal cognitive abilities. When presented with speech in adverse listening conditions, listeners can rely on contextual, semantic and lexical information in order to decode the unfamiliar speech. Evidence suggests that lexical information in particular is important for adapting to unfamiliar speech. For example, listeners adapt to hearing noise-vocoded speech when presented with semantically meaningless (Loebach, Pisoni, & Svirsky, 2010) and syntactically incorrect (Davis et al., 2005) sentences; when listeners are presented with non-words, however, adaptation does not always occur. This has been shown with noise-vocoded speech (Davis et al., 2005; however, see also Hervais-Adelman et al., 2008), and during perceptual adaptation to ambiguous phonemes (Norris et al., 2003). Lexical information likely provides listeners with a framework from which they can estimate the probability of an unfamiliar speech sound belonging to a particular phonemic category (e.g. Lively et al., 1994; Norris et al., 2003); without this framework (for example, in non-words), recognition of the unfamiliar speech becomes more difficult.

Research has also demonstrated a ‘pop-out’ effect during recognition of acoustically degraded speech, whereby providing the listener with a contextual cue leads to much greater recognition accuracy (Giraud et al., 2004), similar to pop-out effects observed in visual perception (Ahissar & Hochstein, 2004). Furthermore, greater perceptual adaptation to noise-vocoded speech can be achieved if listeners undergo training that exploits the pop-out effect by providing feedback after each word (Davis et al., 2005; Hervais-Adelman et al., 2008; Wayne & Johnsrude, 2012); in this way, using

the semantic and lexical cues provided by the feedback allows the listener to more easily learn how to decode the noise-vocoded speech.

As well as contextual cues, there is mounting evidence that recognition of, and perceptual adaptation to, unfamiliar speech relies heavily on cognitive processes. These include executive functions such as attention (Adank & Janse, 2010; Huyck & Johnsrude, 2012), linguistic skills such as lexical knowledge (Borovsky, Elmana, & Fernald, 2012; Janse & Adank, 2012), syntactic processing (Wingfield, McCoy, Peelle, Tun, & Cox, 2006), working memory (Janse & Adank, 2012), and statistical learning (the ability to implicitly detect structural regularities in an input; Neger, Rietveld, & Janse, 2014). Overall cognitive processing speed may also play a part, particularly in older adults who can show signs of 'cognitive slowing'(Janse, 2009). Different cognitive abilities are likely required to process and decode the unfamiliar speech signal, whilst identifying and predicting the lexical and semantic information contained in the speech. Despite the variety of cognitive abilities studied in the literature, we are yet to understand how they interact to ultimately improve a listener's speech recognition. The relationship between different cognitive processes is likely to be highly complex, and variable depending on the listening context and individual. A comprehensive study of a variety of cognitive abilities, and careful consideration of the likely role and contribution of each ability, is required to fully understand how cognition contributes to perceptual plasticity.

Although cognition clearly plays an important role in perceptual plasticity when speech is unfamiliar, lower-level sensory processing is also highly important; for example, in a study of adaptation to noise-vocoded speech, a measure of auditory processing (auditory modulation detection) was the only significant predictor of better adaptation to noise-vocoded speech, when several measures of cognitive function were also taken into account (Erb, Henry, Eisner, & Obleser, 2012). Furthermore, hearing ability is still the largest known predictor of speech recognition and perceptual adaptation in adverse listening conditions (Janse & Adank, 2012). The driving mechanisms of perceptual plasticity therefore likely comprise an interaction between lower-level auditory and higher-level cognitive processes (e.g. Davis & Johnsrude, 2007) – that is, a balance between bottom-up and top-down processing that is weighted according to the listener's exact circumstances and abilities. Indeed, cognitive processes

such as attentional control may in fact modify the use of lower-level acoustic cues and higher-level contextual information during perceptual adaptation (Scharenborg, Weber, & Janse, 2015). However, a comprehensive model of the driving mechanisms of perceptual plasticity in relation to unfamiliar speech processing, is yet to be formed.

1.2.7 Factors affecting recognition of, and adaptation to, unfamiliar speech

Perceptual plasticity of unfamiliar speech may be a robust ability, but it nevertheless varies greatly between individuals and listening contexts. The success of recognition of, or perceptual adaptation to, unfamiliar speech depends on many factors, and these can be categorised in a similar way to adverse listening conditions: 1) factors relating to the speech source and environment, such as the type of speech, or the modality in which it is presented (for example, audiovisual compared to auditory speech); and 2) factors relating to the receiver, such as cognitive ability, or behavioural strategies. In this section, a brief overview of these factors is provided, before focusing on the factors investigated in this thesis: cognitive ability and audiovisual speech.

1.2.7.1 Factors relating to the speech source. Recognition of, and adaptation to, unfamiliar speech, depends greatly on the amount of exposure to the speech variant. For all types of unfamiliar speech, greater exposure leads to greater perceptual adaptation and therefore better speech recognition (Borrie et al., 2012; Clarke & Garrett, 2004; Davis et al., 2005; Mcgarr, 1983; Peelle & Wingfield, 2005). Being exposed to multiple speakers also leads to greater adaptation to accented speech compared to just one speaker (Bradlow & Bent, 2008), as exposure to a greater amount of variation likely helps listeners to constrain possible responses. Although the baseline intelligibility of speech can affect the rate of adaptation (Bradlow & Bent, 2008), a significant amount of adaptation still takes place even when speech is relatively unintelligible (Bradlow & Bent, 2008; Davis et al., 2005).

As discussed in section 1.2.6, providing extra contextual cues such as written feedback can greatly improve adaptation to unfamiliar speech (Borrie et al., 2012; Davis et al., 2005; Loebach et al., 2010), providing compensatory strategies for the listener to exploit when speech is unfamiliar. Similarly, additional perceptual cues can be exploited by the listener when the auditory signal is unclear. For example, being face-to-face with a speaker can provide extra visual cues from facial, head and mouth

movements, which complement or correspond to the auditory signal, to help the listener recognise the unfamiliar speech. The benefits gained from audiovisual speech have been extensively studied in relation to compensating for environmental adverse conditions; however, researchers have now also begun to recognise its potential benefit to recognition of, and adaptation to, unfamiliar speech (for a full discussion on this topic, see section 1.2.10.

1.2.7.2 Factors relating to the listener. Many factors relating to the listener can also influence perceptual plasticity when speech is unfamiliar. Hearing ability contributes greatly to the success of perceptual plasticity of unfamiliar speech recognition; for example, listeners with poorer hearing have been shown to have poorer recognition of accented (Adank & Janse, 2010; Gordon-Salant, Yeni-Komshian, Fitzgibbons, Cohen, & Waldroup, 2013; Janse & Adank, 2012) and time-compressed speech (Gordon-Salant & Fitzgibbons, 1993, 2001; Janse, 2009; Schneider, Daneman, & Murphy, 2005). Increasing age can also make recognition of unfamiliar accented (Adank & Janse, 2010; Gordon-Salant et al., 2010) and time-compressed (Gordon-Salant & Fitzgibbons, 2001; Janse, 2009; Wingfield et al., 2006) speech difficult, even after hearing ability is taken into account. However, perceptual adaptation to unfamiliar speech appears to be relatively preserved in older individuals, with little to no differences in patterns of perceptual adaptation to accented (Gordon-Salant et al., 2010) and time-compressed (Peelle & Wingfield, 2005) speech, although adaptation in older adults may slow more quickly (Adank & Janse, 2010; Peelle & Wingfield, 2005).

As discussed in section 1.2.6, individual differences in cognitive ability are also related to recognition accuracy and perceptual adaptation to unfamiliar speech (this is discussed in more detail in section 1.2.9). Furthermore, individuals with cognitive impairments such as aphasia (language impairments resulting from brain damage; Bruce, To, & Newton, 2012; Newton, Burns, & Bruce, 2013) can have difficulty recognising accented speech. Indeed, age-related declines in cognitive processing may account for some of the age-related differences observed in studies of older and younger adults (Adank & Janse, 2010; Janse, 2009; Mattys & Scharenborg, 2014; Wingfield et al., 2006), although the importance of cognition compared to age-related hearing loss is debated (e.g. Schneider et al., 2005). Nevertheless, investigating the effects of cognitive

ability on individual performance is a growing area of interest in relation to perceptual plasticity in adverse listening conditions.

1.2.8 Summary

To summarise, a wide variety of source-related adverse listening conditions have been studied in the literature, addressing the effects that they have on speech recognition. Listeners are able to adapt rapidly and robustly to these conditions throughout the life span, and this likely results in changes to our perceptual categorisation of phonemes and/or re-tuning of our attention. Evidence shows that these changes are driven by cognitive and sensory processes, but we are still lacking a comprehensive model to fully explain how this occurs. Lastly, myriad factors relating to the speech source and the listener can affect perceptual plasticity; however, much is yet to be understood regarding individual differences in performance. The remainder of this literature review focuses on two specific factors that may go some way to explaining individual differences in recognition of, and perceptual adaptation to, unfamiliar speech: cognitive ability, and audiovisual speech.

1.2.9 Cognitive ability

The role of cognitive ability in recognition of, and perceptual adaptation to unfamiliar speech, has largely been tested in studies of individual differences. A range of abilities have been investigated, for example vocabulary knowledge (Janse & Adank, 2012), attention (Adank & Janse, 2010; Huyck & Johnsrude, 2012), working memory (Janse & Adank, 2012), statistical learning (Neger et al., 2014) and processing speed (Janse, 2009) have all been related to perceptual adaptation to unfamiliar speech. However, the majority of studies have focused largely on working memory processes. Working memory is proposed to be a vital mechanism for recognition of speech in adverse listening conditions (Ronnberg, Rudner, Foo, & Lunner, 2008), as it is likely required to process the unclear speech signal as it unfolds. Indeed, there is a great deal of evidence showing that working memory is important for speech recognition in background noise (see Akeroyd, 2008, for a review). However, there is less evidence that working memory contributes to recognition of source-related adverse conditions, or to perceptual adaptation, as results have largely shown no evidence of a correlation (Ellis & Munro, 2013; Erb et al., 2012; Gordon-Salant et al., 2013; however, see also

Janse and Adank, 2012). Despite the clear contribution of working memory to recognition of speech in background noise, current evidence suggests that it is less likely to be a primary mechanism of perceptual plasticity in relation to unfamiliar speech. However, previous studies have not tested the contribution of working memory in samples large enough to detect small effects (Ellis & Munro, 2013; Erb et al., 2012; Gordon-Salant et al., 2013), and this hypothesis therefore needs to be confirmed in a sufficiently large sample.

Other aspects of cognition, such as linguistic processes, may also be key to recognising and adapting to unfamiliar speech. Evidence has indeed shown the importance of lexical information in perceptual adaptation (e.g. Davis et al., 2005), and greater vocabulary knowledge (as measured by standard vocabulary tests) has predicted greater perceptual adaptation to accented speech in older adults (Janse & Adank, 2012), as well as the prediction of upcoming words in an unfolding passage of speech (Borovsky et al., 2012). If listeners rely on lexical and semantic knowledge when the speech signal is unclear, easily accessing this knowledge may indeed provide a useful compensatory strategy. However, the role of vocabulary knowledge needs to be confirmed in young, normal-hearing adults who are not as dependent on possible compensatory strategies from lexical knowledge.

Executive function has recently been proposed as a key mechanism of perceptual adaptation and recognition of unfamiliar speech. Executive function comprises different components (for example, attention shifting, information monitoring and updating, and inhibition; Miyake et al., 2000). These processes may be crucial to adaptation, for example to focus overall attention on the speech signal, or to tune attention to salient acoustic features (Amitay, 2009). Indeed, attentional control has been related to greater perceptual adaptation to accented speech in older adults (Janse and Adank, 2012). Perceptual adaptation to noise-vocoded speech is also dependent on attention towards the speech signal (Huyck & Johnsrude, 2012; Wild et al., 2012). However, the exact role of executive function in these contexts remains unclear; for example, exactly how attentional abilities help listeners to adapt to unfamiliar speech. Whilst more focused attention on the task or on certain aspects of the stimuli would clearly be beneficial to perceptual adaptation or recognition of unfamiliar speech, it does not explain how ambiguities in unfamiliar speech are resolved, or how the correct

lexical items are selected, for example. Studying particular and separable components of executive function, such as inhibition, may therefore help to constrain theories regarding its role in perceptual plasticity of unfamiliar speech, as well as explaining the mechanisms of perceptual adaptation.

1.2.10 Audiovisual speech

When a listener is face-to-face with a speaker, visual speech information from their mouth, facial and head movements is automatically integrated with the auditory speech signal (McGurk & Macdonald, 1976; Sumby & Pollack, 1954; Summerfield, 1987). In clear listening conditions, this visual speech information is largely redundant; in adverse listening conditions, however, it becomes more salient, and listeners can exploit visual cues to improve their recognition of the speech (Sumby & Pollack, 1954). This 'shift' in attention towards the visual modality is evidenced by a greater visual weighting when listeners are exposed to non-native accented speech (Hazan, Kim, & Chen, 2010), and the observation that listeners with impaired hearing tend to rely more on visual speech cues than listeners with normal hearing (Tye-Murray, Sommers, & Spehar, 2007). Like auditory speech, visual speech can provide both segmental (relating to phonemes) and suprasegmental information (relating to intonation, rhythm and stress patterns; Lansing & McConkie, 1999; Summerfield, 1987; Swerts & Krahmer, 2008). For example, mouth movements can provide information regarding place of articulation for consonants or lip position for vowels, or even stress patterns (Swerts & Krahmer, 2008). Visual speech also provides temporal and spatial information which may help direct the attention of the listener to the speech signal (Grant & Seitz, 2000; Spence, Ranson, & Driver, 2000; Summerfield, 1987).

The benefits of presenting speech in the audiovisual modality are well-known in relation to environmental adverse listening conditions, particularly background noise. Speech recognition is often considerably better in such conditions compared to presentation in the auditory modality (Erber, 1975; Grant, Walden, & Seitz, 1998; A. Macleod & Summerfield, 1987; Ross, Saint-Amour, Leavitt, Javitt, & Foxe, 2007; Sumby & Pollack, 1954). Recent studies have also demonstrated an 'audiovisual benefit' for recognition of unfamiliar speech. Particularly, exposure to audiovisual noise-vocoded speech has led to greater recognition of, and perceptual adaptation to,

noise-vocoded speech (Bernstein, Auer, Eberhardt, & Jiang, 2013; T. Kawase et al., 2009; Pilling & Thomas, 2011). Several studies have also shown that recognition of accented speech is improved by the presence of visual speech cues, both in background noise (Janse & Adank, 2012; S. Kawase, Hannah, & Wang, 2014; Yi, Phelps, Smiljanic, & Chandrasekaran, 2013) and in clear listening conditions (Arnold & Hill, 2001). These results imply that visual speech cues could potentially provide a useful compensatory strategy for perceptual adaptation to unfamiliar speech – that is, audiovisual cues could enhance learning. In a study of older adults, additional visual cues had no overall effect on adaptation to foreign-accented speech in comparison to only auditory cues (Janse & Adank, 2012). However, this could have been due to a confound of age (as older adults are less proficient speech-readers; Sommers, Tye-Murray, & Spehar, 2005) and variable hearing ability, combined with a difficult semantic verification task. The finding therefore needs to be verified in a younger, normal-hearing population to clarify the role of visual speech cues for adaptation to accented speech.

Despite the number of studies investigating audiovisual speech perception, little is known about exactly when listeners make use of visual speech cues in adverse conditions; for example, are visual cues more useful to listeners when speech is first heard, such as at the beginning of a sentence? Similarly, are they more useful to listeners *before* they have adapted to unfamiliar speech, or are they continuously used regardless of any improvements in recognition? Several studies have shown that listeners look more towards a speakers mouth when background noise is present (Buchan, Pare, & Munhall, 2007, 2008; Lansing & McConkie, 2003), and increasingly so as the noise levels increase (Vatikiotis-Bateson, Eigsti, Yano, & Munhall, 1998). Measuring listeners' eye gaze may therefore help to investigate the timing of visual speech processing in adverse listening conditions. Indeed, eye gaze may represent an overt behavioural response by listeners to compensate for unfamiliar or unclear speech. A case study of a single speech-reader assessed eye-tracking as a potential method for use in such contexts, and to understand how eye gaze relates to effective speechreading (Lansing & McConkie, 1994); however, the study has not been followed up. Understanding the mechanisms of eye gaze during audiovisual speech perception could potentially provide a strategy for individuals to improve their recognition of, or perceptual adaptation to, unfamiliar speech. For example, a specific strategy of eye gaze

such as fixating steadily on the speaker's mouth, could potentially help hearing impaired listeners to better recognise audiovisual speech. However, thus far, no study has examined eye gaze during recognition of unfamiliar speech, or during perceptual adaptation.

1.2.11 Summary

Cognitive ability and audiovisual speech cues are potentially important factors affecting recognition of, and adaptation to, unfamiliar speech. Investigating these factors will help to explain individual differences in performance, as well as the mechanisms and behavioural strategies behind adaptation. The cognitive basis of perceptual plasticity of unfamiliar speech is not fully understood, particularly the individual and combined contribution of specific abilities such as executive function, lexical knowledge and working memory. Similarly, the contribution of audiovisual speech cues to perceptual plasticity is not fully known, particularly in relation to adaptation to accented speech, and whether audiovisual cues are exploited by listeners at particular time points during recognition and perceptual adaptation. This thesis therefore presents novel, empirical research investigating these factors in relation to perceptual plasticity of accented and noise-vocoded speech.

As discussed in the introduction, perceptual plasticity (that is, listeners' response to adverse listening conditions) is generally measured in one of two ways: as speech recognition, or as perceptual adaptation. In this review, both types have been discussed as definitions of perceptual plasticity in response to adverse listening conditions. However, in the literature they have largely been investigated separately. These different measurements may rely on different underlying cognitive and neural processes, as well as different behavioural strategies employed by the listener. Nevertheless, there is little empirical evidence which can clarify this distinction, and the measurements are rarely both included in individual studies, thus allowing for any differences to be directly compared (although, see Adank & Janse, 2010; Janse & Adank, 2012). It is therefore useful for any comprehensive investigation into perceptual plasticity of speech under adverse conditions to address both types of perceptual plasticity, and both of these measurements were therefore investigated in each of the experiments presented in this thesis.

1.3 Aims and Hypotheses

Chapter 3. Cognitive predictors of perceptual adaptation to accented speech

Aims. To assess the contribution of cognitive abilities (executive function, vocabulary knowledge and working memory) to recognition of, and adaptation to, accented speech.

Hypotheses. Executive function and vocabulary knowledge will significantly contribute to recognition of, and adaptation to, accented speech, while working memory will contribute to a lesser extent.

Chapter 4. Audiovisual cues benefit recognition of accented speech in noise but not perceptual adaptation

Aims. To test whether perceptual adaptation to accented speech is greater with the presence of audiovisual speech cues (that is, when the listener is face-to-face with the speaker) compared to audio-only speech cues.

Hypotheses. Perceptual adaptation to accented and noise-vocoded speech will be greater with the presence of audiovisual speech cues compared to audio-only speech cues.

Chapter 5. Eye gaze during recognition of audiovisual noise-vocoded speech

Aims. To investigate when listeners use audiovisual speech cues during recognition of, and perceptual adaptation to, audiovisual noise-vocoded speech; specifically, to investigate eye gaze towards a speaker's eyes and mouth during i) recognition of individual sentences, and ii) exposure to multiple sentences (perceptual adaptation). As a secondary aim, the relationship between eye gaze and recognition of audiovisual noise-vocoded speech was also investigated.

Hypotheses. 1) Eye gaze will be increasingly directed towards a speaker's mouth during recognition of individual noise-vocoded sentences; this will peak towards

the start of the sentence, and then decrease towards the end. 2) Eye gaze towards the speaker's mouth will decrease during perceptual adaptation to noise-vocoded speech, as speech recognition improves. 3) Eye gaze towards a speaker's mouth will be related to recognition of audiovisual noise-vocoded speech.

CHAPTER 2

GENERAL METHODS

This chapter summarises the general methodology used in all studies presented in the thesis, as well as some specific methods used in particular studies, such as cognitive tests and eye-tracking. Details of the methods used, as well as justifications for their use are provided. Further details on the methods employed are described in the methods section of each study.

2.1 Participants

In all studies, the participants were young, healthy individuals with normal hearing and normal or corrected-normal vision. Hearing and visual acuity were assessed before each study using standard audiometric and visual tests, and any participants who did not meet the inclusion criteria for these measures (described below) were excluded before taking part. The recruitment criteria required participants to be native British English speakers aged between 18-30 with no history of speech, language or neurological impairments. Potential participants were screened for these criteria and if they did not meet them, they were not included in any of the studies. A young, healthy population was tested in order to gain baseline data without the confounds of age and sensory or cognitive impairment. Participants were all recruited through advertisements at the University of Manchester and comprised both staff members and students from different faculties, although a large number were from the Psychology undergraduate programme. All participants gave their written, informed consent, and no single participant took part in more than one experiment. Participants were given either monetary compensation or course credits for their time. Ethical approval was provided by the University of Manchester Ethics Committee (ref: 11350).

2.1.1 Hearing acuity. Hearing acuity can greatly affect performance in tasks of speech recognition (Akeroyd, 2008) and perceptual adaptation to unfamiliar speech (e.g. Adank & Janse, 2010). Therefore, all our participants had hearing within the normal range for young adults. Before each experiment, participants' hearing was measured using pure-tone audiometry for the main audiometric frequencies in speech (0.5, 1, 2 and 4 kHz) in both ears. Any participant with a hearing threshold level greater than 20

dB for more than one frequency in either ear, was excluded and did not participate in any of the studies. Across all studies, one male and three female participants were excluded based on these criteria.

2.1.2 Visual acuity. The studies reported in chapters 4 and 5 involved perception of visual as well as auditory speech stimuli, and normal or corrected-normal visual acuity was therefore a key criteria for inclusion in these studies. Participants' vision was tested before each experiment, and they were included in the study if their corrected binocular vision was 6/6 or better using a reduced Snellens chart, and their stereoacuity was at least 60 seconds of arc using a TNO test. Across all studies, one male and five females were excluded based on these criteria.

2.2 Materials

2.2.1 Accented speech. Accented speech is the source-related adverse condition that is perhaps most frequently encountered in everyday life. Most adults in the UK, whether through real life or the media, will have experience of adapting to an unfamiliar accent. Furthermore, certain populations have difficulty recognising and adapting to accented speech, such as hearing impaired (Adank & Janse, 2010) or cognitively impaired (Bruce et al., 2012) adults. In addition to these reasons, we particularly chose to study accented speech for the studies reported in chapters 3 and 4, as it is a naturalistic stimulus that demonstrates the natural processes of perceptual adaptation, and represents an ecologically valid stimulus from which to study the mechanisms of perceptual plasticity. Many different types of accented speech have been addressed in the literature. Non-native (e.g. Spanish and Chinese; Bradlow & Bent, 2008), regional (e.g. Northern-Irish English; Pinet et al., 2011) and novel accents (Adank & Janse, 2010; Maye et al., 2008) have all previously been studied in relation to perceptual adaptation.

A common problem with investigating perceptual adaptation to a particular accent is that listeners may have previously been exposed to it, particularly if participants are recruited from a multicultural environment such as a university. As the amount of exposure to a particular accent affects perceptual adaptation, controlling for familiarity is important. One solution is to use a novel accent (Adank & Janse, 2010; Janse & Adank, 2012; Maye et al., 2008) – that is, an accent that has been created

artificially for experimental purposes. Using this type of stimulus to measure speech recognition also allows comparison with a familiar accent produced by the same speaker (Adank & Janse, 2010).

For the studies reported in chapters 3 and 4, a completely novel English accent was created, based on the process used in Adank and Janse (2010) to create a novel Dutch accent. To do this, the vowel patterns of a standard Southern British English accent were systematically varied, by substituting certain vowels, in an iterative process. All vowels were native British English so the resulting accent would be perceived as an unfamiliar regional accent. The accent was developed using an iterative process in order to achieve an accent that was relatively unintelligible, thus leaving room for adaptation, but that was not impossible to understand (for full details, and a phonetic description of the accent, see Chapter 3, Methods).

There are pros and cons to using a novel accent in experimental work. Cristia et al. (2012) question the validity of a novel accent as it is uncertain whether perceptually adapting to a novel accent is the same as adapting to a real one. The novel accent presented in this thesis, for example, only varied in terms of its phonemic patterns and did not include variations in prosodic or stress patterns that may be present in genuine accents. However, a similar argument could also apply to natural accents, as variation in accented speech differs between accents (for example, a Spanish accent differs from a Chinese accent), but also between speakers of the same accent. The particular accent or speaker that is used in any given study, whether real or novel, is therefore unique, and this variation poses a problem for any study of accented speech, and the generalisation of any conclusions drawn from it.

The study in Chapter 3 investigated the cognitive predictors of perceptual adaptation to accented speech, and any differences in familiarity between particular individuals would particularly be a confound, as this would affect their amount and rate of adaptation. It was therefore decided to use a novel accent to ensure that the experiment was well controlled in this respect. The same novel accent was used in Chapter 4 (Study 1) to investigate the use of audiovisual speech cues in adaptation to accented speech. In Study 2 of Chapter 4, a genuine non-native accent was used to verify whether listeners would use audiovisual speech cues more for a non-native accent

compared to a novel one. A Japanese accent (produced by a native Japanese speaker) was chosen, as it was unlikely that this accent was frequently encountered by our participants. To verify this, all participants were additionally asked to rate their familiarity with Japanese accents on a scale of 1-7.

2.2.2 Noise-vocoded speech. The study reported in chapter 5 investigated perceptual adaptation to noise-vocoded speech. Noise-vocoding alters the spectral (frequency) information of the speech that is available to the listener, while retaining temporal cues and changes in amplitude (loudness and softness). It is created by dividing the speech into frequency bands, extracting the amplitude envelope from each frequency band, and then applying the amplitude envelope in each range to band-passed noise (Shannon et al., 1995). The resulting speech sounds like a harsh whisper, but varies in intelligibility depending on how many bands are using to create it; that is, increasing the number of bands also increases the intelligibility of the noise-vocoded speech (Davis et al., 2005; Loizou et al., 2003; Shannon et al., 1995).

Noise-vocoded speech has been quite widely investigated in the perceptual adaptation literature (Davis et al., 2005; Hervais-Adelman et al., 2008; Loebach et al., 2010; Shannon et al., 1995). As it was originally created to simulate hearing with a cochlear implant (Shannon et al., 1995), it is also studied theoretically as an adverse condition relating to the listener (e.g. Faulkner et al., 2000). Noise-vocoded speech was selected as the speech stimulus in Chapter 5 due to evidence that audiovisual speech cues are particularly beneficial to listeners in this context. Particularly, listeners adapt more to noise-vocoded speech with audiovisual speech cues than audio-only cues (Bernstein et al., 2013; T. Kawase et al., 2009; Pilling & Thomas, 2011). For this reason, it would likely provide a reliable stimulus from which to study eye gaze during perceptual adaptation to unfamiliar audiovisual speech.

2.2.3 IEEE sentences. All experiments presented in this thesis used randomly selected sentences from the Institute of Electronics Engineers (IEEE) Harvard sentences (IEEE, 1969) as the testing stimuli (see Appendix B for examples). The IEEE sentences comprise several lists of sentences that were originally developed to test telephone lines, but that have been used extensively in speech perception and audiological research (e.g. Hawley, Litovsky, & Culling, 2004; Loebach et al., 2010; Rosen et al., 2013). They

were selected for use in this thesis due to their relatively consistent length and structure and, in particular, their relative unpredictability in terms of their semantic and linguistic content. They are considered to be slightly more challenging than other sentence batteries such as the BKB sentences, thus providing a stimulus that would likely generate sufficient variation in listeners' responses. The same sentences were used in all experiments to provide consistency, and to control for any effects of stimuli type and length between studies.

2.3 Tests of Cognitive Ability

In Chapter 3, participants' cognitive ability was measured using neuropsychological tests which would subsequently be related to their performance on the speech recognition task. Such tasks are quick and easy to administer, and have been used widely in psychological research. They have been used in many studies of individual differences in speech recognition in adverse conditions (for example, see Akeroyd, 2008), as well as perceptual adaptation (e.g. Adank & Janse, 2010; Janse & Adank, 2012). Some neuropsychological tests are used primarily for clinical diagnoses, for example tests from standard IQ test batteries, while others are developed specifically for research purposes. However, the principal behind all of these tests is that in performing a specific behavioural task, participants are required to use a specific cognitive ability (for example working memory). A problem with this assumption is that such tests are not direct measurements and could therefore tap into multiple cognitive processes, depending on the individual's strategy in carrying out the task, or their level of involvement; for example, all neuropsychological tests require a certain amount of attention and mental focus, and so may inadvertently test an individual's effort, attention or even their overall energy levels as well as the intended cognitive ability.

Nevertheless, neuropsychological tests remain the simplest method for measuring a variety of cognitive abilities, and provide data from which further experimental hypotheses can be created. They also provide a suitable method for studying individual differences in cognitive ability, for example in relation to perceptual adaptation. Neuropsychological tests were therefore used to measure participants' cognitive ability in Chapter 3 in an individual differences analysis – specifically,

working memory, linguistic skills, and executive function. The following neuropsychological tests were selected to measure these abilities.

2.3.1 Executive Function. A standard Stroop test was used to measure executive function, or more specifically, inhibition. The Stroop test was first documented in 1935 (Stroop, 1935) and has been studied extensively in the field of psychology (see C. M. Macleod, 1991 for a review). It involves participants completing three separate but related tasks, each one timed. In the first, participants are requested to name the colour of several rows of blue, red and green squares from left to right, as accurately and as quickly as possible. In the second task, they are required to read written words of the same colours, in the same manner. Finally, the first and second parts are combined so that participants are presented with the words written in incongruous colours (for example the word 'red' written in blue ink). The final task is therefore more difficult and is invariably performed more slowly than either the reading or colour-naming tasks, thus indicating an interference effect from the written words when naming the colours.

The Stroop test is believed to measure inhibitory mechanisms in behaviour (C. M. Macleod, 1991; Miyake et al., 2000). Inhibition is a component of attention that relates to the regulation of dominant, automatic or prepotent behavioural responses (Miyake et al., 2000), often in the presence of conflicting perceptual information. In the Stroop test, participants perceive conflicting semantic and linguistic information, and must inhibit the (dominant) incorrect verbal response which is primed by perceiving the written word. Although related to attention, inhibition has been shown to comprise a separate cognitive component (Miyake et al., 2000), and the Stroop test provides a reliable and robust measurement of this ability, whilst also demonstrating considerable individual variation.

2.3.2 Vocabulary knowledge. Vocabulary knowledge indicates an individual's ability to learn and encode phonetic information, but also to map this to semantic concepts; a test of vocabulary knowledge may measure the strength of these mappings as well as long term memory and retrieval processes for lexical items. Vocabulary knowledge also indicates linguistic experience, as it can increase throughout the lifespan even into the sixtieth decade, when other cognitive abilities start to decline (Schaie,

Willis, & Ohanlon, 1994; Singer, Verhaeghen, Ghisletta, Lindenberger, & Baltes, 2003). A standardised vocabulary test from the Wechsler Abbreviated Scale of Intelligence (WASI; Wechsler, 1999) was used. In this test, participants are presented with individual words of increasing 'difficulty' (that is, words presented later in the test have a lower frequency and more complex meaning; for example, 'panacea' compared to 'bird'). As each word is presented, the participant is asked to describe what the word means as accurately as possible, in their own words. Following the test guidelines, the experimenter can prompt the participant once for more information if they provide a correct but slightly vague definition, by saying "Can you say a bit more?", but no other prompts or feedback are given.

2.3.3 Working memory. To measure working memory, a reading span test was used. The reading span test (Ronnberg, Lyxell, Arlinger, & Kinnefors, 1989) is a computer-based task, adapted into English from the original Swedish version (Daneman & Carpenter, 1980). It was developed specifically for research purposes, and has been particularly used for research into speech recognition in adverse listening conditions (for example, Ellis & Munro, 2013; Lyxell & Holmberg, 2000; Zekveld, Rudner, Johnsrude, & Ronnberg, 2013). The test comprises several stages, each one increasing in difficulty, and testing how well the participant can recall specific lexical items. For each stage, participants are first requested to read aloud a series of sentences presented incrementally (word-by-word) on screen; in the first stage this comprises three sentences, increasing to four, five and then six sentences in later stages. After each sentence, participants are required to state whether the sentence made sense semantically by saying "yes" or "no", for example they might read "The tall tree laughed" to which they should respond "no". After each series of sentences, the participant is then asked to recall either the first or last word from each sentence that they have just heard, in the order in which they heard them. Participants are never told whether they will be asked to recall the first or last word, therefore preventing them from using a particular attentional strategy. Furthermore, a time limit of 20 seconds is imposed for participants to recall all words from each series.

The reading span test is challenging, and few participants are able to recall the maximum number of sentences with perfect accuracy. It provides a measure of working memory that specifically relates to linguistic skills but that forces the participant to

process the linguistic information that they read, as well as holding it in memory and ultimately recalling specific, unpredictable parts of it. This test was selected as it has previously been related to recognition of speech in noise (for reviews see Akeroyd, 2008; Besser, Koelewijn, Zekveld, Kramer, & Festen, 2013), and because it relates to the recall of linguistic information rather than, for example, numerical information as in digit span tasks. The memory processing element also reflects a more ecologically valid test in relation to speech recognition, for example in comparison to recalling a list of digits.

2.4 Experimental Design and Analyses

Two experimental designs were employed for the experiments in this thesis: correlational and mixed design (with between-group and within-participant factors).

2.4.1 Correlational design. A correlational design is used in Chapter 3 to examine the relationship between cognition and perceptual adaptation to accented speech, while in Chapter 5 it is used to examine the relationship between eye gaze and perceptual adaptation to distorted audiovisual speech. This type of design is appropriate for assessing individual differences in speech recognition and perceptual adaptation, and is the most commonly used design to investigate the role of cognition in such studies (Adank & Janse, 2010; Akeroyd, 2008; Erb et al., 2012; Gordon-Salant et al., 2013; Janse & Adank, 2012; Lyxell & Ronnberg, 1989). It has also been used in audiovisual speech perception studies to assess whether measurements of eye gaze are related to particular behaviours such as speech recognition (Buchan et al., 2007; Everdell, Marsh, Yurick, Munhall, & Pare, 2007; Lansing & McConkie, 2003).

Although a correlational design cannot ascertain causation between any two variables, it does indicate the presence of a relationship – for example, which cognitive abilities are involved in adaptation, and whether eye gaze is related to recognition of audiovisual speech. It can therefore still inform us about the mechanisms of perceptual adaptation and recognition of unfamiliar speech. Furthermore, the design lends itself to use of regression models which can indicate the relative contribution of individual predictors (as well as their combined contribution), and also path analysis which can indicate mediation effects between variables. These analyses may be particularly useful

in determining the importance of different cognitive abilities, and constructing a model for cognition during perceptual adaptation and speech recognition.

2.4.2 Mixed experimental design: between-group and within-participant. In Chapters 4 and 5, mixed experimental designs were employed to investigate between-group and within-participants effects, and interactions between them. In both studies, group comparisons assessed differences in speech recognition for audiovisual compared to audio-only speech, while within-participant comparisons assessed improvements in speech recognition over time – that is, perceptual adaptation. Using the modality of speech as the experimental manipulation (and time as a pseudo-manipulation) allowed causation to be inferred, and interactions between modality and improvements over time to be investigated – that is, whether the modality of speech affected speech recognition and the amount of perceptual adaptation. Mixed ANOVAs (as well as non-parametric equivalents where necessary) were thus used for these analyses.

2.5 Speech Recognition Task

In all studies, the same task was used to measure speech recognition. This was a simple repetition task whereby participants were requested to repeat out loud each sentence that they heard, in their own voice (and not imitating the accent, when applicable). They were asked to repeat as much or as little of the sentence as they could, even if this was just one or two words. This repetition task is commonly used in studies of speech recognition and perceptual adaptation (e.g. Adank & Janse, 2010; Peelle & Wingfield, 2005; Rosen et al., 2013) and provides a simple means of assessing how well an individual has heard and recognised speech items. Although it is not necessarily ecologically valid (repeating what one has just heard is not a usual occurrence in everyday conversation), it does provide an accurate measurement of recognition accuracy. Participants were scored on the percentage of keywords (content and function words) that they correctly repeated in each sentence. In this way, the task better reflected how we process heard speech in an everyday setting, as listeners likely focus their attention on key words that convey the gist of what is being said, rather than on, for example, filler words which do not carry as much meaning.

2.6 Eye-Tracking

Chapter 5 used eye-tracking to record participants' eye movements while perceiving audiovisual unfamiliar speech. The study of eye movements dates back as far as the late 1800's (Javal, 1878), but in the last 30 years eye movements have been extensively researched in relation to cognitive and perceptual processes, most likely due to advances in eye tracking technology (see Kowler, 2011 for a review). Modern eye trackers use an infra-red camera, and sometimes an illuminator, to track an individual's pupil and corneal reflection. Eye trackers can be mounted onto an individual's head while they view the real world, or they can be placed on a desk in front of an individual while they look at a screen. Numerous aspects of eye gaze can be recorded for analysis, but the most common measurements in psychological research are saccades (rapid eye movements from one spatial location to another), and fixations (periods of time between saccades when the eyes are relatively stationary). These can be analysed both temporally and spatially, as eye trackers have both a high temporal and spatial resolution.

Psychological studies have examined eye gaze in relation to a variety of perceptual contexts, such as visual search and reading (see Liversedge & Findlay, 2000; Rayner, 1998, for reviews). Recordings of eye gaze have also been used to investigate face perception (for a review, see Birmingham & Kingstone, 2009; Yarbus, 1967) and, to a lesser extent, audiovisual speech perception (e.g. Buchan et al., 2007; Vo, Smith, Mital, & Henderson, 2012). The advantages of using eye tracking in psychological research is that eye gaze can indicate certain cognitive processes relating to the allocation of visual attention, that are otherwise not easily measurable. For example, visual world paradigms (whereby participants are presented with pictures while listening to a passage of speech) have been used extensively in psycholinguistic research to understand how the meaning and syntax of sentences is processed (see Huettig, Rommers, & Meyer, 2011 for a review). Similarly, reading studies have been able to measure the timing and sequence of eye movements to understand how they relate to word identification or syntactic parsing, for example (Rayner & Reichle, 2010).

In relation to the research presented in this thesis, eye gaze may inform us as to when individuals gain and use visual linguistic information from a speaker's face. Of

particular interest is the contrast between looking at a speaker's eyes, most likely for social reasons (Birmingham & Kingstone, 2009; Langton, Watt, & Bruce, 2000), and looking at their mouth to gain useful linguistic cues in adverse listening conditions (Vatikiotis-Bateson et al., 1998). To this end, Chapter 5 used eye tracking to measure patterns of eye gaze during recognition of unfamiliar audiovisual speech. Specifically, a desk-mounted Eyelink 1000 (SR Research, Ontario, Canada) was used to track participants' eye movements (that is, the pupil and corneal reflection), while they viewed videos or static images of a speaker on a screen in front of them. Although this is a somewhat unnatural set-up, desktop eye-trackers are commonly used to measure eye gaze in an experimental setting and allow for the presentation of audiovisual stimuli, whilst controlling for head and body movements, and the extent of the visual stimuli (that is, only eye movements on the screen and thus on the visual stimulus are recorded).

Using eye gaze to infer cognitive processes has some limitations, primarily that it relies on the assumption that attention is related to foveal (central) vision, rather than parafoveal or peripheral vision (Liversedge & Findlay, 2000). Furthermore, it also assumes that attention is directed towards the visual modality. This may be a particular problem in using eye gaze to study audiovisual speech, as weighting of auditory and visual speech information can vary (Hazan et al., 2010). Nevertheless, it provides a relatively unexplored and interesting method with which to investigate audiovisual speech perception, and has never been used to investigate perceptual adaptation to unfamiliar speech. Particularly, it will allow an investigation of whether eye gaze plays a role in recognition of, and perceptual adaptation to, unfamiliar speech.

CHAPTER 3

COGNITIVE PREDICTORS OF PERCEPTUAL ADAPTATION TO ACCENTED SPEECH

Manuscript published in the Journal of the Acoustical Society of America, April 2015

Cognitive predictors of perceptual adaptation to accented speech

Briony Banks[a)]
School of Psychological Sciences, University of Manchester, Manchester M13 9PL, United Kingdom

Emma Gowen
Faculty of Life Sciences, University of Manchester, Manchester M13 9PL, United Kingdom

Kevin J. Munro and Patti Adank[b)]
School of Psychological Sciences, University of Manchester, Manchester M13 9PL, United Kingdom

(Received 8 July 2014; revised 19 February 2015; accepted 2 March 2015)

The present study investigated the effects of inhibition, vocabulary knowledge, and working memory on perceptual adaptation to accented speech. One hundred young, normal-hearing adults listened to sentences spoken in a constructed, unfamiliar accent presented in speech-shaped background noise. Speech Reception Thresholds (SRTs) corresponding to 50% speech recognition accuracy provided a measurement of adaptation to the accented speech. Stroop, vocabulary knowledge, and working memory tests were performed to measure cognitive ability. Participants adapted to the unfamiliar accent as revealed by a decrease in SRTs over time. Better inhibition (lower Stroop scores) predicted greater and faster adaptation to the unfamiliar accent. Vocabulary knowledge predicted better recognition of the unfamiliar accent, while working memory had a smaller, indirect effect on speech recognition mediated by vocabulary score. Results support a top-down model for successful adaptation to, and recognition of, accented speech; they add to recent theories that allocate a prominent role for executive function to effective speech comprehension in adverse listening conditions. © *2015 Acoustical Society of America*. [http://dx.doi.org/10.1121/1.4916265]

[CGC] Pages: 2015–2024

I. INTRODUCTION

The ability to recognize speech in adverse listening conditions is a robust and flexible mechanism that is supported by our ability to "tune in" to unfamiliar or distorted speech (for reviews, see Samuel and Kraljic, 2009; Cristia *et al.*, 2012; Mattys *et al.*, 2012). Such perceptual adaptation can be defined as improved speech recognition (that is, accessing the semantic content of the speech message through perceiving the acoustic signal) as a result of exposure to an unfamiliar speech type. Despite the robustness of this ability, the relative success of perceptual adaptation can vary, and may depend on individual differences in the cognitive ability of the listener.

While it is increasingly acknowledged that certain cognitive abilities (such as working memory or executive function) play an important role in perceptual adaptation to unfamiliar speech (Adank and Janse, 2010; Erb *et al.*, 2012; Huyck and Johnsrude, 2012; Janse and Adank, 2012), no comprehensive model exists to explain the cognitive mechanisms underlying this ability. Given that adapting to adverse listening conditions is an inherent part of human communication, understanding the mechanisms underlying perceptual adaptation will contribute to existing models of speech recognition as well as a growing body of research into communication in adverse conditions, which is relevant to both healthy and clinical populations.

The role of cognition has been widely investigated in relation to auditory processing in normal-hearing and hearing-impaired populations (e.g., Pichora-Fuller and Singh, 2006), particularly for recognition of speech in noise (for a review, see Akeroyd, 2008). However, it is not known whether such findings translate to perceptual adaptation to unfamiliar speech, particularly in a young, normal-hearing population. Existing accounts of speech perception currently emphasize the role of working memory in optimal and adverse listening conditions; for example, the ease of language understanding model (Ronnberg *et al.*, 2008) proposes that in difficult conditions, memory *storage* is required to keep track of the unfolding speech signal, while memory *processing* is required when speech input does not match existing phonological representations. Although working memory is a relatively reliable predictor for recognition of speech-in-noise (for normal-hearing and hearing-impaired adults; Akeroyd, 2008), evidence for a strong relationship between working memory and adaptation to unfamiliar speech is limited. Janse and Adank (2012) observed a relationship between working memory and recognition of a novel accent; however, this has not been replicated for perception of non-native (Gordon-Salant *et al.*, 2013), frequency compressed (Ellis and Munro, 2013) or noise-vocoded (Erb *et al.*, 2012) speech. There are three possible explanations for this limited evidence. First, it could be that working memory does not play as prominent a role in perceptual adaptation to unfamiliar speech as predicted by the ease of language understanding model; indeed, the model

[a)]Author to whom correspondence should be addressed. Electronic mail: briony.banks@manchester.ac.uk
[b)]Current address: Division of Psychology and Language Sciences, University College London, Chandler House, 2 Wakefield Street, London WC1N 1PF, United Kingdom.

endeavors to predict ease of understanding rather than speech recognition *per se* (Ronnberg, 2003). Second, the effect of working memory may be relatively subtle and the aforementioned studies may not have had the required statistical power to detect a small effect. Third, perceptual adaptation to unfamiliar speech may be primarily driven by other cognitive abilities (such as executive function or linguistic abilities) while working memory may have a more indirect influence similar to that observed for speech reading (Lyxell and Ronnberg, 1989), or for perceptual adaptation to degraded visual input (Kennedy *et al.*, 2009).

Behavioral and neuroimaging research has indeed provided support for a role of executive function during perceptual adaptation to unfamiliar speech. Executive function has been defined as cognitive processes, such as inhibitory mechanisms, that help control and coordinate other aspects of cognition, and is associated with activity in the frontal lobe (e.g., Miyake *et al.*, 2000). Neuroimaging studies have revealed activity in cortical regions associated with executive function when processing degraded compared with clear speech (Wild *et al.*, 2012; Erb *et al.*, 2013), while behavioral studies have demonstrated that attentional mechanisms are recruited for perceptual adaptation in lower level auditory training (Halliday *et al.*, 2011), and higher level adaptation to noise-vocoded (Huyck and Johnsrude, 2012), frequency-compressed (Ellis and Munro, 2013), and accented speech (Adank and Janse, 2010; Janse and Adank, 2012). However, it is unclear exactly how executive functions contribute to perceptual adaptation. Attentional control may certainly aid the listener to direct attention to the more salient aspects of the perceived speech (Amitay, 2009), or to better attend to the cognitively demanding input. Nevertheless, this does not explain how perceivers are able to learn and adapt to the new speech patterns of an unfamiliar accent, particularly how perceptual ambiguities are resolved or how correct lexical items are identified and selected. Successful perceptual adaptation may therefore be supported by inhibitory processes that facilitate the identification of correct lexical items and inhibit incorrect responses. Although measures of inhibition have predicted successful speech recognition in noise (Sommers and Danielson, 1999; Janse, 2012; Koelewijn *et al.*, 2012), they have thus far not been related to perceptual adaptation to unfamiliar speech.

Linguistic abilities, and particularly processing of lexical information, may also contribute to perceptual adaptation to unfamiliar speech. Studies have demonstrated that the lexical positioning of ambiguous phonemes affects subsequent perceptual categorisation of that phoneme (Norris *et al.*, 2003; Eisner and McQueen, 2005) and that intact lexical information is important for adaptation to noise-vocoded speech (Davis *et al.*, 2005). Nevertheless, only one study to date has investigated individual vocabulary knowledge as a predictor of perceptual adaptation to unfamiliar speech; in a study of older adults, Janse and Adank (2012) observed that better vocabulary knowledge predicted greater adaptation to accented speech. Given that vocabulary knowledge is relatively preserved in an older population, particularly in comparison to working memory and executive function (Schaie *et al.*, 1994; Singer *et al.*, 2003), a reliance on vocabulary knowledge in this population may reflect a compensatory strategy rather than the normal route to adaptation in younger adults. To confirm whether this finding generalizes to a wider population, it is therefore necessary to also test a younger, normal-hearing population as a baseline measure.

Given the evidence described above, we propose that inhibition and vocabulary knowledge substantially contribute to perceptual adaptation to unfamiliar speech, while working memory contributes to a lesser extent. These three abilities have not previously been tested together in a single model of perceptual adaptation, thus allowing for their relative individual importance, as well as their combined contribution, to be examined. Testing these abilities in a large sample from a young, healthy population will enable detection of smaller effects while controlling for confounding factors of age-related sensory and cognitive decline. Furthermore, previous research has either focused on overall recognition of unfamiliar speech, or on adaptation (improvement in recognition accuracy) over time; we propose that these measures may tap into different cognitive processes and that both should be included in studies of speech perception in adverse listening conditions. The present study therefore investigated the contribution of three cognitive abilities (inhibition, vocabulary knowledge, and working memory) in adaptation to, and recognition of, accented speech. We chose to investigate accented speech as it is a naturalistic variant that is pertinent to everyday communication and, although adaptation to other distortions (such as noise-vocoded speech) likely involve the same mechanisms, it is not known whether they can be directly compared. We tested younger adults to build on previous results from older adults while providing baseline evidence from a cognitively healthy and normal-hearing population. Our hypothesis was that better abilities in the three cognitive measures would lead to greater and more rapid adaptation and to better overall recognition accuracy of the accented speech, with inhibition and vocabulary knowledge accounting for a greater amount of variance than working memory.

II. METHOD

A. Participants

One hundred students (24 male; mean age, 20.4 years; standard deviation, 2.28; range 18–30 years) recruited from the University of Manchester, participated in the study (for a linear multiple regression analysis with four predictor variables, a sample size >95 is required to detect an effect size of 0.15 [$\alpha = 0.05$, $1 - \beta = 0.85$], Faul *et al.*, 2009). All participants were native British English speakers with no history of neurological, psychiatric, speech, or language problems (self-declared). Participants' hearing was assessed using pure-tone audiometry at 0.5, 1, 2, and 4 kHz in each ear separately. Any participant with a hearing threshold level >20 dB for more than one frequency in either ear was excluded from the study. We provided compensation of course credit or £7.50 for participation. The study was approved by The University of Manchester ethics committee, and all participants gave their written informed consent.

B. Materials

Stimulus material consisted of 105 Institute of Electrical and Electronics Engineers (IEEE) Harvard sentences (IEEE, 1969), selected because of their low predictability and standardized structure and length. We transcribed 90 of the sentences into a novel accent (Maye *et al.*, 2008; Adank and Janse, 2010). We chose to use a novel accent as a naturalistic stimulus that avoids confounds from participant familiarity and allows for a matched-guise design (Lambert *et al.*, 1960); that is, we could create stimuli from the same speaker in a standard and novel accent. The accent was created by systematically changing the vowel sounds of a standard British English accent, using vowel sounds from a variety of English regional accents (e.g., Scottish, Irish and Northern English; see Table I for the full phonetic transcription). This was achieved through an iterative process where we maintained the length of the vowel sounds (long, short or diphthongs) so as not to affect stress patterns. Our aim was to create an accent that would be unfamiliar to all participants but also of relatively low intelligibility (in order to measure adaptation over time, we required an accent with low intelligibility to avoid ceiling effects in earlier trials); to this end, some vowels sounds were not modified at all (that is, they remained as standard British English vowels). When asked about the accent after the experiment, the majority of participants indicated that it "sounded a bit like" an existing regional English accent (e.g., Scottish or Irish) but could not identify it.

A 30-year-old male speaker with a Standard British English accent was trained in the novel accent to provide all accented stimuli for the experiment. Recordings were made in a sound-treated laboratory with a SM58 microphone (Shure Inc., Niles, IL). All recordings were manually checked by the experimenter for pronunciation accuracy and naturalness, and any that were not deemed suitable (e.g., due to mispronunciation) were excluded from the study. Ninety novel accented sentences were divided into 6 lists of 15 sentences to be used as the testing stimuli. A further 15 sentences recorded by the same speaker in a Standard British English accent were selected to be the baseline "unaccented" sentences (see Sec. II C for details). All audio files were normalized by equating the root-mean-square amplitude, resampled at 22 kHz in mono (over both ears) and cropped at the nearest zero crossings at voice onset and offset, using Praat software (Boersma and Weenink, 2012).

TABLE I. Phonetic description of the novel accent.

International Phonetic Alphabet	Example
ɪ → ɛ	sit → set
ɛ → ɪ	bet → bit
æ → ɛ	hat → het
ʌ → ʊ	cud → could
ɜː → ɛə	girl → gairl
aː → ɔː	dark → dork
ɒ → ɔː	hot → hawt
ɔː	door
uː	food
ʊ	good
ə	mother
iː	tree
ɛə → ɜː	hair → her
əʊ → aʊ	vote → vowt
aʊ → uː	how → hoo
ɛɪ → aɪ	way → wye
aɪ → ɔɪ	my → moy
ɪə	hear
ɔɪ	joy

C. Procedure

Participants wore sound attenuating headphones (HD 25-SP II; Sennheiser electronic GmbH & Co. KG, Wedemark, Germany) for the duration of the experiment. The volume level was adjusted to a comfortable level by the experimenter for the first participant and then kept at the same level for all participants thereafter. Stimuli were presented using MATLAB software (R2010a, MathWorks, Natick, MA; see Sec. II E for full details). To familiarize participants with the procedure, and to gain a baseline measurement of recognition accuracy for native speech, participants first listened to the 15 unaccented sentences as practice trials, followed by the 90 accented sentences. Sentence lists were counterbalanced across the six testing blocks, each comprising 15 sentences, and were presented in a pseudo-random order per testing block and per participant. Each sentence was presented once to each participant to avoid training effects of particular items. Last, participants were tested on the three cognitive measures. The experiment was carried out in one session lasting approximately 60 min. As part of a wider study, participants also underwent training with additional versions (audiovisual, audio-only or visual-only) of the novel-accented stimuli between block 3 and block 4; however, no significant effects of training were observed,[1] and these results will not be discussed further in this paper.

D. Speech recognition task

After presentation of each sentence, we instructed participants to repeat out loud as much or as little of the sentence as they could, in their normal voice and without imitating the accent. The experimenter scored participants' responses immediately after each trial according to how many keywords out of a possible four were correctly repeated. These responses were logged using MATLAB to determine the signal-to-noise ratio (SNR) of the next trial (see Sec. II E for details). No feedback was given to participants. Keywords comprised either content or function words and, in line with previous studies of perceptual adaptation to unfamiliar speech (Dupoux and Green, 1997; Golomb *et al.*, 2007), were marked as correct despite incorrect suffixes (such as -s, -ed, -ing) or verb endings. If only part of a word (including compound words) was repeated it was counted as incorrect. If a participant repeated a word imitating the novel accent (that is, if their pronunciation deviated from their own accent to match the novel accent), this was also counted as incorrect, as we could not ascertain whether the participant had correctly identified the lexical item, or whether they had simply repeated the phonological pattern they had heard.

E. Speech reception thresholds

Recognition accuracy during each testing block was measured by establishing participants' Speech Reception Thresholds (SRTs) in speech-shaped background noise, using an adaptive staircase procedure (Plomp and Mimpen, 1979). Measuring speech recognition in this way avoids ceiling effects associated with rapid perceptual adaptation to accented speech, and also controls for variation in individual baseline comprehension. Accuracy (number of correctly repeated keywords) was maintained at 50% by adjusting the SNR in pre-determined steps. Thus, as perceptual adaptation took place and correct responses increased, the SNR was decreased and the task became increasingly difficult (Baker and Rosen, 2001). The procedure was carried out using MATLAB software. The initial SNR for the first sentence in each block was 10 dB. Throughout the staircase procedure, the background noise varied in steps of 8 dB for the first two reversals, and 2 dB for each reversal thereafter. The mean SNR for all reversals per testing block indicated the SRT measurement for each participant.

F. Cognitive background measures

Vocabulary knowledge was tested using the Wechsler Abbreviated Scale of Intelligence (WASI; Wechsler, 1999) vocabulary subtest, which requires participants to provide oral definitions of words. Participants were scored according to the standard instructions, and overall percentages were calculated for analysis. Inhibition was measured using a standard Stroop test (Stroop, 1935), presented to the participant on paper and requiring oral responses. The test comprised three sections: Color naming (C), word naming (W), and word-color interference (WC), whereby participants were required to name the (incongruent) color of the ink that words were written in. Each section was timed manually by the experimenter using a stopwatch. Interference scores, based on the mean time (in seconds) to complete each section, were calculated using the following equation:

$$\text{Interference} = \text{WC} - (\text{W} \times \text{C})/(\text{W} + \text{C}).$$

Finally, working memory was tested using an English version of a standard reading span test (Ronnberg *et al.*, 1989). This requires participants to read 3–6 sentences which appear on screen word-by-word, and then to subsequently recall either the first or last word of each sentence when prompted by the experimenter. The total number of correctly recalled words was calculated for analysis.

G. Data analysis

Within our data set, we identified two outliers (one for the accented SRTs and one for the unaccented SRTs) with standardized residuals >3.29, and these scores were modified to the value of the group mean SRT plus two standard deviations. Interference scores for the Stroop test were positively skewed, so the data were log transformed to allow for parametrical analysis. Mauchly's test indicated that the assumption of sphericity had been violated for the repeated measures analysis of variance (ANOVA), $\chi^2(14) = 75.61$, $p < 0.001$, therefore degrees of freedom were corrected using Huynh-Feldt estimates of sphericity ($\varepsilon = 0.86$). Unless otherwise stated, all other assumptions for parametrical testing of the data were met.

Recognition of unfamiliar speech can be measured in two ways: As overall performance, or as improvement in performance over time, and both of these measures were used in our analyses of individual differences. Overall performance (recognition accuracy) was calculated as the mean SRT across all testing blocks. Adaptation was analyzed as the amount and rate of improvement. We calculated the amount of adaptation as the difference in mean SRTs between the first three and the last three testing blocks, while rate of adaptation was calculated by fitting a linear function to the recognition accuracy data (Erb *et al.*, 2012); we used the equation $y = mx + b$, where y is the mean SRT, x is time (block), m is the slope, and b is the intercept. The slope of each participant's linear fit was used as a measurement of adaptation rate. To investigate individual differences in perceptual adaptation, we used multiple linear regression to analyze the relationships between recognition accuracy, amount and rate of adaptation (our dependent variables), and four predictor variables: unaccented SRTs (representing participants' baseline ability to deal with speech in noise), and vocabulary, working memory and Stroop interference scores. We included unaccented SRTs in order to examine relationships between the cognitive predictors and comprehension when unaccented SRTs were held constant; that is, we could infer that the individual contribution of each cognitive measure was related to the accented speech over and above the background noise.

To test our hypothesis that working memory may have an indirect effect on comprehension (that is, that the relationship between working memory and comprehension was mediated by other predictors), we used path analysis, fitting a hypothesized model to our data and thus assessing the direct and indirect (mediated) effects between variables. Model fit was assessed using the chi-square (χ^2) statistic, the root-mean-square error of approximation (RMSEA), and the Tucker-Lewis Index (TLI). As our sample size was relatively small for this type of analysis, we used bootstrapping (Shrout and Bolger, 2002; Preacher and Hayes, 2004) to construct bias-corrected confidence intervals (95%) to test for mediation effects between variables.

III. RESULTS

A. Perceptual adaptation to accented speech

Table II shows the mean SRTs for the unaccented speech (hereafter "unaccented"), as well as SRTs for each testing block of the accented speech. As SRTs represent the signal-to-noise ratio (dB), higher levels reflect poorer tolerance to background noise (poorer performance). As expected, unaccented SRTs were significantly lower than mean accented SRTs for each testing block, even after correcting for multiple comparisons [block 1, $t(99) = -21.20$, $p < 0.001$; block 2, $t(99) = -21.68$, $p < 0.001$; block 3, $t(99) = -14.76$, $p < 0.001$; block 4, $t(99) = -14.18$, $p < 0.001$; block 5,

TABLE II. Mean SRTs and standard deviations per testing block.

Testing block	Mean (dB)	Standard deviation (dB)
Unaccented	0.57	1.7
1	10.18	4.54
2	7.54	3.16
3	5.48	3.50
4	6.54	4.42
5	5.60	2.95
6	4.03	2.71

$t(99) = -15.45$, $p < 0.001$; block 6, $t(99) = -13.14$, $p < 0.001$], confirming that the novel accent negatively affected participants' performance. To confirm whether participants' tolerance to background noise significantly changed over time, we carried out a repeated-measures ANOVA to examine within-subject effects of testing block (6 levels). We observed a significant main effect of testing block [$F(4.32, 409.88) = 45.72$, $p < 0.001$, $\eta_p^2 = 0.33$]. Pairwise comparisons (Bonferroni correction, $p < 0.003$) revealed that SRTs for blocks 2, 3, 4, 5 and 6 were significantly lower than SRTs in block 1, confirming that participants' tolerance to the background noise increased. SRTs for blocks 3, 5 and 6 were also significantly lower than block 2, and block 6 was significantly lower than blocks 3, 4 and 5. As expected, there was considerable individual variation between participants' SRTs throughout the experiment (see Fig. 1). There was a significant negative correlation between the slope and intercept of all linear fits ($r = -0.79$, $p < 0.001$), indicating that participants who initially performed worse improved the most.[2]

B. Cognitive ability and perceptual adaptation to accented speech

Table III shows the correlation matrix between adaptation amount, adaptation rate and recognition accuracy for the accented speech, and the four predictor variables.

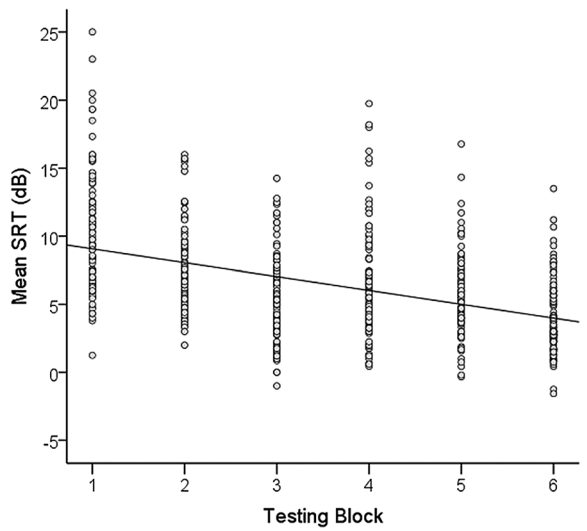


FIG. 1. Individual variation in recognition accuracy of accented speech in noise: Mean SRTs (in dB) per participant, per testing block, with mean linear fit for all participants.

Adaptation amount was negatively correlated with Stroop scores ($r = -0.29$, $p = 0.004$; see Fig. 2), indicating that lower interference scores (and thus better inhibition) was related to greater adaptation. Adaptation rate (slope) was positively correlated with Stroop scores ($r = 0.21$, $p = 0.04$, indicating that better inhibition was related to a faster rate of adaptation (it should be noted that, as lower SRTs indicated better performance, adaptation slopes had mainly negative values ($M = -1.01$); lower values of our adaptation rate measurement therefore represent faster adaptation). Recognition accuracy was positively correlated with unaccented SRTs ($r = 0.36$, $p < 0.001$), indicating that participants who could tolerate a high level of background noise for the unaccented sentences, could also tolerate a high level of background noise for the accented sentences. Recognition accuracy was negatively correlated with vocabulary ($r = -0.38$, $p < 0.001$) and working memory ($r = -0.22$, $p = 0.03$); that is, participants with better vocabulary and working memory scores had lower SRTs, and thus had better recognition accuracy of the accented speech. Between the four predictor variables, working memory was positively correlated with vocabulary (better working memory was related to greater vocabulary knowledge, $r = 0.25$, $p = 0.01$), and negatively correlated with Stroop interference scores (better working memory was related to greater inhibition, $r = -0.25$, $p = 0.01$). Vocabulary was negatively correlated with unaccented SRTs (greater vocabulary knowledge was related to better recognition accuracy of the unaccented sentences, $r = -0.39$, $p < 0.001$). Between the three outcome variables, recognition accuracy and adaptation rate were negatively correlated, $r = -0.23$, $p = 0.01$ (participants with poorer overall recognition accuracy adapted more quickly), and adaptation amount and rate were negatively correlated, $r = -0.84$, $p < 0.001$ (participants who adapted the most did so at a faster rate). No issues of collinearity were identified, and thus, all cognitive measures and the unaccented SRTs could be included in our regression analyses.

In order to analyze the contribution of the four predictor variables to recognition accuracy, adaptation amount and adaptation rate, we carried out three backward stepwise regression analyses. Table IV shows the results of the regression model for recognition accuracy of the accented speech. When all other predictor variables were held constant, unaccented SRTs ($\beta = 0.27$, $p = 0.008$) and vocabulary ($\beta = -0.24$, $p = 0.02$) significantly predicted recognition accuracy, whereas working memory did not ($\beta = -0.16$, $p = 0.09$). Table V shows the results of the regression models for adaptation amount and adaptation rate. In both models, Stroop scores (inhibition) significantly predicted the amount ($\beta = -0.29$, $p = 0.004$) and rate ($\beta = 0.21$, $p = 0.04$) of adaptation.

As we had observed a significant correlation between working memory and recognition accuracy, but working memory did not significantly predict recognition accuracy in our regression model, we hypothesized that there was an indirect relationship between these two variables, mediated by vocabulary score. We carried out a path analysis to test this hypothesis. The presence of correlations between the three variables (working memory, vocabulary and

TABLE III. Correlation matrix for recognition accuracy of, and adaptation to, accented speech and cognitive ability, with means and standard deviations ($N = 100$).[a]

Variable	Mean	Standard deviation	Recognition	Adaptation amount	Adaptation rate	Unaccented	Vocabulary	Working memory	Stroop
Recognition Accuracy (SRT, dB)	6.50	2.05	_						
Adaptation amount (dB)	2.34	2.18	0.03	_					
Adaptation rate (slope)	−1.01	0.68	−0.23[b]	−0.84[c]	_				
Unaccented (SRT, dB)	0.57	1.70	0.36[c]	0.03	−0.03	_			
Vocabulary (%)	66.76	6.88	−0.38[c]	0.07	−0.04	−0.39[c]	_		
Working Memory (%)	49.76	9.06	−0.22[b]	0.13	−0.08	−0.01	0.25[b]	_	
Stroop	1.52	0.09	0.15	−0.29[d]	0.21[b]	−0.09	−0.14	−0.25[b]	_

[a]Higher mean scores for recognition accuracy and Stroop indicate poorer performance. Higher scores for all other variables indicate better performance.
[b]Two-tailed Pearson's correlations, significant at $p < 0.05$.
[c]Two-tailed Pearson's correlations, significant at $p < 0.001$.
[d]Two-tailed Pearson's correlations, significant at $p < 0.01$.

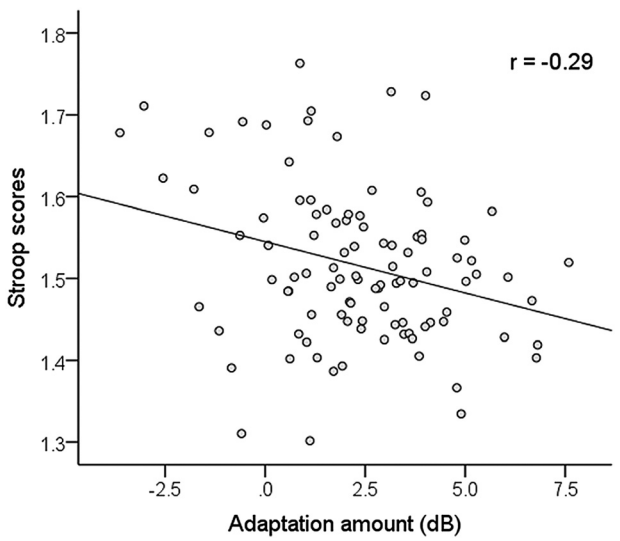


FIG. 2. Scatterplot showing correlation between amount of adaptation to accented speech and Stroop interference scores (inhibition), with linear regression best fit; r = correlation coefficient.

TABLE IV. Backward stepwise regression analysis for the predictors of recognition accuracy of accented speech ($N = 100$).[a]

Variable	B	Standard error B	β
Step 1			
Unaccented SRTs	0.35	0.12	0.28[b]
Vocabulary	−0.07	0.03	−0.22[c]
Working memory	−0.03	0.02	−0.13
Stroop	2.50	2.04	0.12
Step 2			
Unaccented SRTs	0.33	0.12	0.27[b]
Vocabulary	−0.07	0.03	−0.24[c]
Working memory	−0.04	0.02	−0.16

[a]$R^2 = 0.24$ for Step1; $\Delta R^2 = -0.01$ for Step 2 ($p < 0.05$).
[b]$p < 0.01$.
[c]$p < 0.05$.

recognition accuracy), meant that our data met the assumptions required for a mediation effect (Baron and Kenny, 1986). It should be noted that these assumptions were not met for the predictors of adaptation amount or rate, and so

TABLE V. Backward stepwise regression analysis for the predictors of (a) amount of adaptation and (b) rate of adaptation (slope) to accented speech ($N = 100$).

Variable	B	Standard error B	β
V(a) Adaptation amount[a]			
Step 1			
Unaccented	0.02	0.14	0.01
Vocabulary	0.01	0.04	0.03
Working memory	0.01	0.03	0.05
Stroop	−6.22	2.36	−0.27
Step 2			
Vocabulary	0.01	0.03	0.02
Working memory	0.01	0.03	0.06
Stroop	−6.26	2.32	−0.27
Step 3			
Working memory	0.01	0.02	0.06
Stroop	−6.30	2.30	−0.27
Step 4			
Stroop	−6.64	2.22	−0.29[b]
V(b) Adaptation rate[c]			
Step 1			
Unaccented	−0.01	0.05	−0.02
Vocabulary	0.00	0.01	0.01
Working memory	−0.01	0.01	−0.03
Stroop	1.40	0.76	0.19
Step 2			
Unaccented	−0.01	0.04	−0.01
Working memory	−0.01	0.01	−0.04
Stroop	1.40	0.75	0.20
Step 3			
Working memory	−0.01	0.01	−0.04
Stroop	1.41	0.74	0.20
Step 4			
Stroop	1.47	0.71	0.21[c]

[a]$R^2 = 0.09$ for Step 1; $\Delta R^2 = 0.00$ for Steps 2, 3 and 4 (p's < 0.05).
[b]$p < 0.05$.
[c]$R^2 = 0.04$ for Step 1; $\Delta R^2 = 0.00$ for Steps 2, 3 and 4 (p's < 0.05).

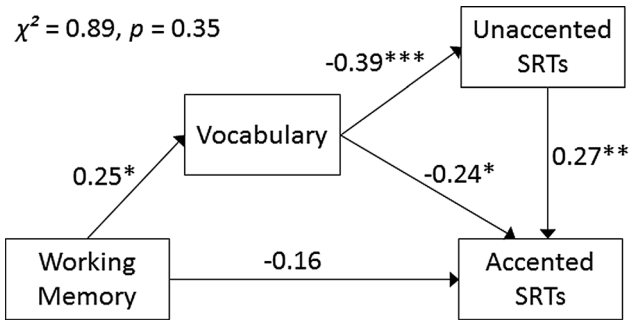


FIG. 3. Path analysis model for the cognitive predictors of recognition accuracy of accented speech. All path parameters are standardized coefficients (direct effects). χ^2 = chi-square statistic (non-significant value indicates the model is a good fit). The pathway between working memory and accented SRTs was not significant ($p > 0.05$) and was mediated by vocabulary score. There was an indirect effect of working memory on accented SRTs, $\beta = -0.09$, $p < 0.01$, and an indirect effect of vocabulary score on accented SRTs, $\beta = -0.11$, $p < 0.01$. * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$.

path analyses to test for mediation effects were not carried out on these data. Figure 3 shows the path model for the predictors of recognition accuracy with standardized coefficients. The inclusion of each pathway was based on observations from our data, while the direction of each pathway was based on our hypotheses (e.g., that vocabulary score predicted recognition accuracy). The model fit the data well: $\chi^2(1) = 0.89$, $p = 0.35$; TLI = 1.02; RMSEA < 0.001. As predicted, the relationship between working memory and recognition accuracy of the accented speech was mediated by vocabulary score; that is, working memory had an indirect effect on recognition accuracy, $\beta = -0.09$, $p < 0.01$, via vocabulary score. Vocabulary had a direct effect on recognition accuracy, $\beta = -0.24$, $p < 0.01$, and an indirect effect on recognition accuracy, $\beta = -0.11$, $p < 0.01$, via unaccented SRTs; vocabulary therefore accounted for the greatest amount of total variance (combined direct and indirect effects) on recognition accuracy, $\beta = -0.34$, $p < 0.01$.

IV. DISCUSSION

The present study investigated how individual differences in cognitive ability relate to perceptual adaptation to accented speech, as measured by overall performance (recognition accuracy) and amount of improvement (adaptation). We predicted that better inhibition (a measure of executive function) and vocabulary knowledge, supported by better working memory, would lead to better recognition accuracy and greater adaptation.

A. Perceptual adaptation to accented speech

As predicted from previous studies of adaptation to accented speech (Clarke and Garrett, 2004; Bradlow and Bent, 2008; Maye *et al.*, 2008; Adank and Janse, 2010; Gordon-Salant *et al.*, 2010; Janse and Adank, 2012), we observed significant improvements in recognition accuracy of our novel accent over time, represented by a greater tolerance to background noise in later compared to earlier trials. As expected, we observed considerable individual variation in SRTs throughout all testing blocks, and participants who had poorer starting levels adapted the most. Similar adaptation patterns have been observed for comprehension of noise-vocoded speech (Stacey and Summerfield, 2007; Erb *et al.*, 2012).

Adaptation to accented speech can occur rapidly, even after as few as eight sentences (Clarke and Garrett, 2004). However, by using a relatively difficult novel accent and an adaptive procedure to vary the background and target SNR, this process was slowed; indeed, our participants continued to improve significantly until the final block of stimuli, after exposure to 90 sentences. The disadvantage of this procedure is that the measure of recognition accuracy obtained (SRTs) represents responses to the accented speech and to the background noise. Although we cannot completely separate both elements, several factors provide evidence that listeners adapted predominantly to the accent, and not to the background noise. First, mean SRTs for the accented speech were significantly different to SRTs for the unaccented speech; that is, participants never perceived the accented speech as well as the unaccented speech, even after exposure to all 90 test sentences. Second, Adank and Janse (2010) demonstrated that SRTs while listening to a standard native accent (using the same adaptive procedure as in the present study) remain stable in a young population, with a difference of <1 dB in SRTs after exposure to 60 sentences. Third, neither of our adaptation measures was significantly correlated with unaccented SRTs, indicating that the amount and rate participants adapted was not related to their ability to process unaccented speech in background noise. This supports our claim that the adaptation we observed in our study (a mean improvement of 6 dB between the first and final testing blocks) was likely related to the accent rather than to the background noise. However, one further limitation should be acknowledged—that the perception of the same speaker with an unfamiliar accent, after listening to him speak with a standard British English accent, may have influenced the higher SRTs in the first block.

B. Cognitive ability and perceptual adaptation to accented speech

Our analyses revealed that inhibition, as measured by the Stroop test, predicted adaptation to the accented speech. Participants who had better inhibition (that is, performed better at the Stroop test) adapted more and at a faster rate than participants who demonstrated poorer inhibition, thus supporting our hypothesis. To our knowledge, ours is the first study to directly link inhibition to perceptual adaptation to accented speech. This finding adds to a growing body of evidence that executive function, such as inhibition or attention, has a major role in perceptual adaptation to unfamiliar speech (Huyck and Johnsrude, 2012; Wild *et al.*, 2012; Erb *et al.*, 2013), including adaptation to accented speech (Adank and Janse, 2010; Janse and Adank, 2012). Inhibitory abilities are likely recruited when competing (and incorrect) lexical responses are triggered by the accented speech (Brouwer *et al.*, 2012; Tuinman *et al.*, 2012), thus helping to resolve ambiguities in the speech signal. This may allow the listener to identify the correct lexical items and thus match unfamiliar phonemic patterns to existing phonemic

representations, resulting in adaption to the patterns of the accented speech. Greater inhibitory abilities may thus allow listeners to overcome ambiguous or unfamiliar auditory input such as accented speech.

Performance on the Stroop test has also been linked to recognition of speech in background noise in older adults (Sommers and Danielson, 1999; Janse, 2012). As our participants listened to the accented speech in background noise, this may explain part of the relationship between Stroop scores and adaptation observed in our study. However, if this were the case, we would also expect the Stroop scores and our adaptation measures to correlate with SRTs for the unaccented speech. No such correlations were observed, which indicates that the relationship between Stroop scores and adaptation reflects efficient adaptation to the accent rather than to the background noise. Nevertheless, it should be noted that our participants only listened to 15 unaccented sentences—fewer than in previous studies that observed a relationship between Stroop scores and speech recognition in noise (Sommers and Danielson, 1999; Janse, 2012); therefore, we may not have observed a correlation between unaccented SRTs and Stroop scores due to the small amount of exposure. A third possible interpretation of our findings is that the Stroop test relates to more than one aspect of executive function, or to individual strategies such as attention or motivation. Although it is not possible to separate the cognitive constructs of the Stroop test in this experiment, overall strategies such as motivation or attention would likely apply to all three cognitive predictors, whereas only Stroop scores were significantly related to adaptation.

Our second finding was that vocabulary knowledge predicted recognition accuracy of the accented speech. As we hypothesized, participants who had greater vocabulary scores could tolerate more background noise overall, and thus their recognition of the accented speech was more robust than participants with lower vocabulary scores. This confirms a role for vocabulary knowledge during perception of accented speech in a young, healthy population, and supports similar findings in older adults (Janse and Adank, 2012). Our path analysis revealed that vocabulary knowledge accounted for the greatest amount of total variance in recognition of the accented speech. We observed a direct relationship between vocabulary knowledge and recognition of the accented speech, but we also observed an indirect relationship via recognition of the unaccented speech (that is, unaccented SRTs partially mediated the relationship between vocabulary score and accented SRTs). Vocabulary score also fully mediated the relationship between working memory and recognition of the accented speech. This suggests a particular importance for lexical knowledge in successfully perceiving native and non-native speech in noise. Greater vocabulary knowledge likely allows the listener to more readily identify and access lexical items from unfamiliar or ambiguous auditory input; stronger mapping between lexical and semantic representations may also help listeners to process the incremental speech input by helping them to anticipate upcoming words in the sentence (Borovsky *et al.*, 2012). Although the role of lexical processing in perceptual adaptation to other speech distortions is debated, for example, noise-vocoded (Hervais-Adelman *et al.*, 2008) and time-compressed (Janse, 2009) speech, lexical information may be particularly pertinent to comprehension of accented speech (e.g., Norris *et al.*, 2003), perhaps aiding the listener to identify patterns of phonetic variation by allowing them to map this variation more easily onto lexical items. However, a second interpretation of our finding is also possible. Vocabulary knowledge is usually correlated with verbal and non-verbal IQ (Wechsler, 1958; Kamphaus, 2005), and indeed, the test used in our study is part of a standard IQ test battery. Our findings here may thus reflect a relationship between speech recognition and general intelligence, rather than specifically with vocabulary knowledge, although measures of IQ have not consistently been found to predict recognition of native speech in noise (Akeroyd, 2008). As we did not test our participants' full IQ, further investigation is required to confirm whether lexical knowledge in particular, or general intelligence, are important for successful recognition of accented speech.

Vocabulary knowledge did not predict amount or rate of adaptation to the accented speech as we had hypothesized, which is contrary to results observed in older adults (Janse and Adank, 2012). These discrepant findings may reflect differences in the populations tested; as vocabulary knowledge can increase into the sixth decade (Schaie *et al.*, 1994) and remains relatively stable into the eighth (Singer *et al.*, 2003), it may provide an important compensatory strategy in older adults following a decline in other cognitive functions.

The third cognitive ability we investigated was working memory. Although we observed a significant correlation between working memory and recognition accuracy, this ability did not directly predict recognition accuracy or adaptation when unaccented SRTs and vocabulary score were also included in our regression analysis. However, working memory did have an indirect relationship with recognition accuracy, mediated by vocabulary knowledge, in our path analysis model. Working memory may therefore support recognition of accented speech via other cognitive abilities (in this case, vocabulary knowledge), as observed in speech reading (Lyxell and Ronnberg, 1989) and perceptual adaptation to distorted visual input (Kennedy *et al.*, 2009). Other studies investigating working memory and perceptual adaptation to unfamiliar speech have produced mixed results: although working memory is the most reliable predictor of recognition of speech in background noise, this is not a wholly consistent finding (Akeroyd, 2008), and indeed we did not observe a correlation between working memory and unaccented SRTs in our study. Janse and Adank (2012) found that working memory predicts overall recognition accuracy of novel-accented speech in older adults (possibly reflecting greater individual variation in an older population), but no other study has observed this, in foreign-accented (Gordon-Salant *et al.*, 2013), frequency compressed (Ellis and Munro, 2013), or noise-vocoded (Erb *et al.*, 2012) speech.

Our findings, together with current evidence, suggest therefore that working memory does not always play a prominent role in perceptual adaptation to, or recognition of, unfamiliar speech. Furthermore, our effects were small even in a

sample of 100 participants. Studies with smaller samples, and particularly in a young, clinically normal population, may therefore be underpowered to detect such small effects. However, another explanation is also possible. The working memory test used in this study (Ronnberg *et al.*, 1989) relies specifically on lexical recall, and responses are scored as incorrect if participants recall the correct semantic concept, but not the exact lexical item (e.g., "gun" instead of "pistol"). An overlap with the abilities required for the vocabulary knowledge test (that is, robust mapping between lexical items and semantic concepts) could therefore account for the mediation effect observed in our data.

The present study measured two important aspects of perceptual adaptation to accented speech—recognition accuracy and adaptation (that is, overall performance and changes in performance over time). The results from our regression analyses suggest that different cognitive abilities are involved in these different aspects of adaptation (executive function for amount and rate of adaptation; vocabulary knowledge and, to a lesser extent, working memory, for recognition accuracy). Nevertheless, it should be noted that our measures of recognition accuracy and adaptation rate were significantly correlated, and so differences between these two measures should be interpreted with caution. However, no such correlation was observed between recognition accuracy and adaptation amount, and so we can assume that these measures do indeed reflect different abilities.

V. CONCLUSION

The present study evaluated the contribution of cognitive ability to perceptual adaptation to accented speech. Results suggest a prominent role for inhibition in perceptual adaptation, and for vocabulary knowledge in overall recognition accuracy. Recognition accuracy was indirectly supported by working memory, via vocabulary knowledge, which suggests that working memory may play a less prominent role in successful recognition of accented speech. Our study is the first to relate inhibition to perceptual adaptation to unfamiliar speech, and substantiates existing evidence that top-down processing, particularly executive function, is important for adapting to speech in adverse listening conditions. However, further investigations may help to discern the exact role of executive function and vocabulary knowledge in perceptual adaptation to accented speech.

ACKNOWLEDGMENTS

This work was funded by a Biotechnology and Biological Sciences Research Council research studentship awarded to B.B. and by The University of Manchester. The authors thank Stuart Rosen for supplying the adaptive staircase program and the speech-shaped noise.

[1]A one-way ANOVA (5 levels: training group) of SRTs for the accented speech revealed no effect of training group: $F(4, 95) = 0.23$, $p = 0.92$, $\eta_p^2 = 0.01$.

[2]It should be noted that starting values did not significantly contribute to the variance in adaptation when included in our regression models: adaptation amount: $\beta = 0.09$, $t = 0.86$, $p = 0.391$; adaptation rate: $\beta = -0.11$, $t = 1.05$, $p = 0.298$.

Adank, P., and Janse, E. (**2010**). "Comprehension of a novel accent by young and older listeners," Psychol. Aging **25**, 736–740.
Akeroyd, M. A. (**2008**). "Are individual differences in speech reception related to individual differences in cognitive ability? A survey of twenty experimental studies with normal and hearing-impaired adults," Int. J. Audiol. **47**, S53–S71.
Amitay, S. (**2009**). "Forward and reverse hierarchies in auditory perceptual learning," Learn. Percept. **1**, 59–68.
Baker, R., and Rosen, S. (**2001**). "Evaluation of maximum-likelihood threshold estimation with tone-in-noise masking," Br. J. Audiol. **35**, 43–52.
Baron, R. M., and Kenny, D. A. (**1986**). "The moderator mediator variable distinction in social psychological-research—Conceptual, strategic, and statistical considerations," J. Pers. Soc. Psychol. **51**, 1173–1182.
Boersma, P., and Weenink, D. (**2012**). "Praat: Doing phonetics by computer (version 5.3.05) [computer program]," www.praat.org (Last viewed February 19, 2012).
Borovsky, A., Elmana, J. L., and Fernald, A. (**2012**). "Knowing a lot for one's age: Vocabulary skill and not age is associated with anticipatory incremental sentence interpretation in children and adults," J. Exp. Child Psychol. **112**, 417–436.
Bradlow, A. R., and Bent, T. (**2008**). "Perceptual adaptation to non-native speech," Cognition **106**, 707–729.
Brouwer, S., Mitterer, H., and Huettig, F. (**2012**). "Can hearing puter activate pupil? Phonological competition and the processing of reduced spoken words in spontaneous conversations," Q. J. Exp. Psychol. **65**, 2193–2220.
Clarke, C. M., and Garrett, M. F. (**2004**). "Rapid adaptation to foreign-accented English," J. Acoust. Soc. Am. **116**, 3647–3658.
Cristia, A., Seidl, A., Vaughn, C., Schmale, R., Bradlow, A. R., and Floccia, C. (**2012**). "Linguistic processing of accented speech across the lifespan," Front. Psychol. **3**, 479.
Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., and McGettigan, C. (**2005**). "Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences," J. Exp. Psychol. Gen. **134**, 222–241.
Dupoux, E., and Green, K. (**1997**). "Perceptual adjustment to highly compressed speech: Effects of talker and rate changes," J. Exp. Psychol. Hum. Percept. Perform. **23**, 914–927.
Eisner, F., and McQueen, J. M. (**2005**). "The specificity of perceptual learning in speech processing," Percept. Psychophys. **67**, 224–238.
Ellis, R. J., and Munro, K. J. (**2013**). "Does cognitive function predict frequency compressed speech recognition in listeners with normal hearing and normal cognition?," Int. J. Audiol. **52**, 14–22.
Erb, J., Henry, M. J., Eisner, F., and Obleser, J. (**2012**). "Auditory skills and brain morphology predict individual differences in adaptation to degraded speech," Neuropsychologia **50**, 2154–2164.
Erb, J., Henry, M. J., Eisner, F., and Obleser, J. (**2013**). "The brain dynamics of rapid perceptual adaptation to adverse listening conditions," J. Neurosci. **33**, 10688–10697.
Faul, F., Erdfelder, E., Buchner, A., and Lang, A.-G. (**2009**). "Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses," Behav. Res. Methods **41**, 1149–1160.
Golomb, J. D., Peelle, J. E., and Wingfield, A. (**2007**). "Effects of stimulus variability and adult aging on adaptation to time-compressed speech," J. Acoust. Soc. Am. **121**, 1701–1708.
Gordon-Salant, S., Yeni-Komshian, G. H., Fitzgibbons, P. J., Cohen, J. I., and Waldroup, C. (**2013**). "Recognition of accented and unaccented speech in different maskers by younger and older listeners," J. Acoust. Soc. Am. **134**, 618–627.
Gordon-Salant, S., Yeni-Komshian, G. H., Fitzgibbons, P. J., and Schurman, J. (**2010**). "Short-term adaptation to accented English by younger and older adults," J. Acoust. Soc. Am. **128**, EL200–EL204.
Halliday, L. F., Moore, D. R., Taylor, J. L., and Amitay, S. (**2011**). "Dimension-specific attention directs learning and listening on auditory training tasks," Att. Percept. Psychophys. **73**, 1329–1335.
Hervais-Adelman, A., Davis, M. H., Johnsrude, I. S., and Carlyon, R. P. (**2008**). "Perceptual learning of noise vocoded words: Effects of feedback and lexicality," J. Exp. Psychol. Hum. Percept. Perform. **34**, 460–474.

Huyck, J. J., and Johnsrude, I. S. (**2012**). "Rapid perceptual learning of noise-vocoded speech requires attention," J. Acoust. Soc. Am. **131**, EL236–EL242.
Institute of Electrical and Electronics Engineers (**1969**). "IEEE recommended practices for speech quality measurements," IEEE Trans. Aud. Electroacoust. **17**, 227–246.
Janse, E. (**2009**). "Processing of fast speech by elderly listeners," J. Acoust. Soc. Am. **125**, 2361–2373.
Janse, E. (**2012**). "A non-auditory measure of interference predicts distraction by competing speech in older adults," Aging Neuropsychol. Cog. J. Norm. Dysfunct. Dev. **19**, 741–758.
Janse, E., and Adank, P. (**2012**). "Predicting foreign-accent adaptation in older adults," Q. J. Exp. Psychol. **65**, 1563–1585.
Kamphaus, R. W. (**2005**). *Clinical Assessment of Child and Adolescent Intelligence*, 2nd ed. (Springer Science, New York), pp. 359–417.
Kennedy, K. M., Rodrigue, K. M., Head, D., Gunning-Dixon, F., and Raz, N. (**2009**). "Neuroanatomical and cognitive mediators of age-related differences in perceptual priming and learning," Neuropsychology **23**, 475–491.
Koelewijn, T., Zekveld, A. A., Festen, J. M., Ronnberg, J., and Kramer, S. E. (**2012**). "Processing load induced by informational masking is related to linguistic abilities," Int. J. Otolaryngol. **2012**, 865731.
Lambert, W. E., Hodgson, R. C., Gardner, R. C., and Fillenbaum, S. (**1960**). "Evaluational reactions to spoken languages," J. Abnorm. Soc. Psychol. **60**, 44–51.
Lyxell, B., and Ronnberg, J. (**1989**). "Information-processing skill and speech-reading," Br. J. Audiol. **23**, 339–348.
Mattys, S. L., Davis, M. H., Bradlow, A. R., and Scott, S. K. (**2012**). "Speech recognition in adverse conditions: A review," Language Cognit. Processes **27**, 953–978.
Maye, J., Aslin, R. N., and Tanenhaus, M. K. (**2008**). "The weckud wetch of the wast: Lexical adaptation to a novel accent," Cognit. Sci. **32**, 543–562.
Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., and Wager, T. D. (**2000**). "The unity and diversity of executive functions and their contributions to complex 'frontal lobe' tasks: A latent variable analysis," Cognit. Psychol. **41**, 49–100.
Norris, D., McQueen, J. M., and Cutler, A. (**2003**). "Perceptual learning in speech," Cognit. Psychol. **47**, 204–238.
Pichora-Fuller, M. K., and Singh, G. (**2006**). "Effects of age on auditory and cognitive processing: Implications for hearing aid fitting and audiologic rehabilitation," Trends Amplif. **10**, 29–59.
Plomp, R., and Mimpen, A. M. (**1979**). "Speech-reception threshold for sentences as a function of age and noise-level," J. Acoust. Soc. Am. **66**, 1333–1342.
Preacher, K. J., and Hayes, A. F. (**2004**). "SPSS and SAS procedures for estimating indirect effects in simple mediation models," Behav. Res. Methods Instrum. Comput. **36**, 717–731.
Ronnberg, J. (**2003**). "Cognition in the hearing impaired and deaf as a bridge between signal and dialogue: A framework and a model," Int. J. Audiol. **42**, S68–S76.
Ronnberg, J., Lyxell, B., Arlinger, S., and Kinnefors, C. (**1989**). "Visual evoked-potentials: Relation to adult speechreading and cognitive function," J. Speech Hear. Res. **32**, 725–735.
Ronnberg, J., Rudner, M., Foo, C., and Lunner, T. (**2008**). "Cognition counts: A working memory system for ease of language understanding (ELU)," Int. J. Audiol. **47**, S99–S105.
Samuel, A. G., and Kraljic, T. (**2009**). "Perceptual learning for speech," Att. Percept. Psychophys. **71**, 1207–1218.
Schaie, K. W., Willis, S. L., and O'Hanlon, A. M. (**1994**). "Perceived intellectual performance change over seven years," J. Gerontol. **49**, P108–P118.
Shrout, P. E., and Bolger, N. (**2002**). "Mediation in experimental and nonexperimental studies: New procedures and recommendations," Psychol. Methods **7**, 422–445.
Singer, T., Verhaeghen, P., Ghisletta, P., Lindenberger, U., and Baltes, P. B. (**2003**). "The fate of cognition in very old age: Six-year longitudinal findings in the Berlin Aging Study (BASE)," Psychol. Aging **18**, 318–331.
Sommers, M. S., and Danielson, S. M. (**1999**). "Inhibitory processes and spoken word recognition in young and older adults: The interaction of lexical competition and semantic context," Psychol. Aging **14**, 458–472.
Stacey, P. C., and Summerfield, A. Q. (**2007**). "Effectiveness of computer-based auditory training in improving the perception of noise-vocoded speech," J. Acoust. Soc. Am. **121**, 2923–2935.
Stroop, J. R. (**1935**). "Studies of interference in serial verbal reactions," J. Exp. Psychol. **18**, 643–662.
Tuinman, A., Mitterer, H., and Cutler, A. (**2012**). "Resolving ambiguity in familiar and unfamiliar casual speech," J. Mem. Lang. **66**, 530–544.
Wechsler, D. (**1958**). *The Measurement and Appraisal of Adult Intelligence*, 4th ed. (The Williams and Wilkins Company, Baltimore, MD), pp. 61–117.
Wechsler, D. (**1999**). *Wechsler Abbreviated Scale of Intelligence* (Pearson, San Antonio, TX), pp. 54–86.
Wild, C. J., Yusuf, A., Wilson, D. E., Peelle, J. E., Davis, M. H., and Johnsrude, I. S. (**2012**). "Effortful listening: The processing of degraded speech depends critically on attention," J. Neurosci. **32**, 14010–14021.

CHAPTER 4.

AUDIOVISUAL CUES BENEFIT RECOGNITION OF ACCENTED SPEECH IN NOISE BUT NOT PERCEPTUAL ADAPTATION

Manuscript published in Frontiers in Human Neuroscience, August 2015

frontiers in Human Neuroscience

ORIGINAL RESEARCH
published: 03 August 2015
doi: 10.3389/fnhum.2015.00422

Audiovisual cues benefit recognition of accented speech in noise but not perceptual adaptation

Briony Banks[1], Emma Gowen[2], Kevin J. Munro[1] and Patti Adank[3]*

[1] School of Psychological Sciences, University of Manchester, Manchester, UK, [2] Faculty of Life Sciences, University of Manchester, Manchester, UK, [3] Speech, Hearing and Phonetic Sciences, University College London, London, UK

Perceptual adaptation allows humans to recognize different varieties of accented speech. We investigated whether perceptual adaptation to accented speech is facilitated if listeners can see a speaker's facial and mouth movements. In Study 1, participants listened to sentences in a novel accent and underwent a period of training with audiovisual or audio-only speech cues, presented in quiet or in background noise. A control group also underwent training with visual-only (speech-reading) cues. We observed no significant difference in perceptual adaptation between any of the groups. To address a number of remaining questions, we carried out a second study using a different accent, speaker and experimental design, in which participants listened to sentences in a non-native (Japanese) accent with audiovisual or audio-only cues, without separate training. Participants' eye gaze was recorded to verify that they looked at the speaker's face during audiovisual trials. Recognition accuracy was significantly better for audiovisual than for audio-only stimuli; however, no statistical difference in perceptual adaptation was observed between the two modalities. Furthermore, Bayesian analysis suggested that the data supported the null hypothesis. Our results suggest that although the availability of visual speech cues may be immediately beneficial for recognition of unfamiliar accented speech in noise, it does not improve perceptual adaptation.

Keywords: speech perception, perceptual adaptation, accented speech, audiovisual speech, multisensory perception

OPEN ACCESS

Edited by:
Ignacio Moreno-Torres, University of Málaga, Spain

Reviewed by:
Jeroen Stekelenburg, Tilburg University, Netherlands
Matthias J. Sjerps, University of California, Berkeley, USA and Radboud University Nijmegen, Netherlands

****Correspondence:***
Briony Banks, School of Psychological Sciences, University of Manchester, 3rd Floor, Zochonis Building, Brunswick Street, Manchester M13 9LP, UK
briony.banks@manchester.ac.uk

Received: *31 March 2015*
Accepted: *10 July 2015*
Published: *03 August 2015*

Citation:
Banks B, Gowen E, Munro KJ and Adank P (2015) Audiovisual cues benefit recognition of accented speech in noise but not perceptual adaptation. Front. Hum. Neurosci. 9:422. doi: 10.3389/fnhum.2015.00422

Introduction

When we encounter a speaker with an unfamiliar accent, we are able to 'tune in' to the new phonetic patterns of speech to understand what they are saying. This type of perceptual adaptation is regularly encountered in daily life and allows us to recognize speech in a variety of native and non-native accents (Clarke and Garrett, 2004; Bradlow and Bent, 2008; Maye et al., 2008). It is a robust ability that is present in all stages of life (for a review, see Cristia et al., 2012) and occurs even with relatively unintelligible accents, albeit at a slower rate (Bradlow and Bent, 2008). The relative success and speed of perceptual adaptation depends on external factors such as the amount and variety of exposure to the accent (Bradlow and Bent, 2008). However, less is known about how the modality of speech can influence the adaptation process – for example, whether adaptation to accented speech is greater when audiovisual speech cues are available, compared to only auditory speech cues. Identifying ways to improve or facilitate this process may benefit communication in certain populations who have difficulty adapting to accented speech, such as older adults

(Adank and Janse, 2010), individuals with aphasia (Bruce et al., 2012), or non-native speakers (Munro and Derwing, 1995); for example, audiovisual speech could be incorporated into language-learning tools or rehabilitation therapies for aphasia.

Perceptual adaptation to accented speech can be seen as a three-stage process: the listener first perceives the new, unfamiliar input; secondly, maps this onto stored lexical items, and thirdly, generalizes these new mappings to other lexical items. Indeed, research has successfully shown that this type of adaptation involves the modification of perceptual phonemic boundaries in relation to perceived lexical items (Norris et al., 2003; Kraljic and Samuel, 2005, 2006); for example, listeners who perceive an ambiguous sound midway between /d/ and /t/ spoken within the word 'croco<u>d</u>ile,' are more likely to then categorize the same sound as /d/ when heard in isolation.

An improvement in perceptual adaptation to accented speech could potentially be achieved by influencing any one of the three stages involved, for example, the first stage may be facilitated through the availability of audiovisual (multisensory) cues. The integration of multisensory input across different sensory modalities can facilitate perception (Stein and Meredith, 1993); for example, auditory perception of speech is improved when integrated with visual input from a speaker's facial movements. Indeed, being face-to-face with a speaker improves speech recognition in noisy environments (Sumby and Pollack, 1954; Erber, 1975; MacLeod and Summerfield, 1987; Grant et al., 1998; Ross et al., 2007), particularly when speech is non-native (Reisberg et al., 1987; Arnold and Hill, 2001; Hazan et al., 2006). Research has shown that audiovisual speech cues help listeners to identify fricative consonants (Jongman et al., 2003) and prosodic cues such as lexical prominence (Swerts and Krahmer, 2008). The benefits of audiovisual cues may also extend to accented speech, as several studies have shown that recognition of accented speech is better for audiovisual compared to audio-only input (Arnold and Hill, 2001; Janse and Adank, 2012; Yi et al., 2013; Kawase et al., 2014). The integration of auditory and visual cues may benefit recognition of accented speech by helping listeners to resolve the perceptual ambiguities of an unfamiliar accent; for example, if a speaker's pronunciation of a particular phoneme or word is unclear, observing their mouth movements may help to identify the correct item. Indeed, exposure to ambiguous audiovisual cues using McGurk stimuli has been shown to influence subsequent phoneme categorization (Bertelson et al., 2003; Vroomen et al., 2004). A listener who is face-to-face with an accented speaker may therefore be able to exploit the perceptual benefit from additional visual input, and adapt more successfully to the accented speech – that is, their recognition of the speech may improve more greatly over time.

Although a large part of everyday communication is carried out face-to-face, most experimental work on accent perception is carried out in the auditory modality, and the use of visual speech information has gained relatively little attention in relation to perceptual adaptation to accented speech. Furthermore, much of the work regarding the potential benefits of audiovisual speech to perceptual adaptation has been carried out using noise-vocoded speech rather than accented. While both speech types are less intelligible than familiar speech, and listeners adapt to them both, variation in noise-vocoded speech stems from degrading the acoustical composition of the entire speech signal, whereas accented speech varies in terms of its phonemic patterns, is acoustically intact and only affects certain speech sounds. Although audiovisual cues have been shown to benefit perceptual adaptation to noise-vocoded speech (Kawase et al., 2009; Pilling and Thomas, 2011; Wayne and Johnsrude, 2012; Bernstein et al., 2013), the observed effects are relatively small and, furthermore, we do not know if such results generalize to accented speech.

Two previous studies have investigated the role of audiovisual cues in perceptual adaptation to accented speech. In a phoneme-recognition study, Hazan et al. (2005) demonstrated that long-term perception of individual non-native phonemes improved when listeners were exposed to audiovisual input, compared to audio-only input; however, this finding was not tested with longer items such as sentences, and it is thus unclear if the results can be generalized to non-native speech in general. Indeed, when Janse and Adank (2012) compared perceptual adaptation to unfamiliar, accented sentences with or without visual cues, they observed no difference in the amount of adaptation, although a small, non-significant trend of greater adaptation during the early stages was present for audiovisual speech. However, two confounding factors may have influenced their findings. The experiment was carried out on older adults, a population that can have particular difficulty with processing visual speech (Sommers et al., 2005); this factor, combined with a relatively difficult semantic verification task, may have rendered the task cognitively demanding for the older participants and negatively affected their performance. Two possible conclusions can therefore be drawn from the two studies described here: first, audiovisual speech cues are not beneficial to perceptual adaptation to longer items of accented speech, although they may improve learning of particular phonemes in isolation (as shown by Hazan et al., 2005); or, audiovisual speech cues do benefit perceptual adaptation to accented speech, but the confounding factors outlined above prevented this effect from being observed. Therefore, evidence from young, healthy adults, using whole sentences and a simple speech recognition task, may help to establish the possible benefits of audiovisual speech cues for perceptual adaptation to accented speech.

We investigated whether audiovisual speech cues do indeed facilitate perceptual adaptation to accented speech. We did this across two studies, each using a different accent and speaker and a different experimental design, but with the same sentences and task. In particular, Study 2 addresses a number of questions arising from Study 1 (see Discussion, Study 1 for details). Study 1 employed a training design similar to those used in studies of noise-vocoded speech (Kawase et al., 2009; Pilling and Thomas, 2011; Wayne and Johnsrude, 2012), and a novel accent to control for familiarity effects (Maye et al., 2008; Adank and Janse, 2010; Janse and Adank, 2012). Participants underwent training in the novel accent with audiovisual or audio-only stimuli, with or without background noise. A visual-only (speech-reading) training condition provided a control group; that is, we did not expect visual training to affect adaptation to the accented

speech. For the pre- and post-training sessions, we presented our accented stimuli in background noise to avoid ceiling effects associated with rapid perceptual adaptation (Janse and Adank, 2012; Yi et al., 2013). We also included two training conditions with background noise for two reasons: firstly, the learning context can influence the outcome of learning (Godden and Baddeley, 1975; Polyn et al., 2009), and consistency between the training and subsequent testing sessions may therefore affect adaptation. As the stimuli in our pre- and post-training sessions were always presented in the context of background noise, we predicted that training with background noise would facilitate recognition of the accented speech in noise following the training. Secondly, we predicted that altering the clarity of the auditory signal (by adding background noise) would increase the use of visual cues during the training (cf. Sumby and Pollack, 1954), and that this would, in turn, increase subsequent adaptation.

If audiovisual cues are beneficial to perceptual adaptation to accented speech, we expected to observe the following: (1) greater adaptation after audiovisual training compared to audio-only or visual-only training; (2) greater adaptation after audiovisual training with background noise compared to audiovisual training in quiet; (3) a greater 'audiovisual benefit' (the difference in adaptation between audiovisual and audio-only training) for the groups trained with background noise, compared to the groups trained without background noise; (4) greater adaptation following all types of training in comparison to visual training (that is, we expected the visual training to have no effect on subsequent recognition of the accented speech). Based on previous evidence that audiovisual cues can benefit recognition of accented speech compared to audio-only cues (Arnold and Hill, 2001; Janse and Adank, 2012; Yi et al., 2013; Kawase et al., 2014), we also expected to observe the following during the training session: (1) better recognition of the accented training stimuli for both audiovisual groups compared to the audio-only groups; (2) poorer recognition of the training stimuli presented in background noise compared to quiet; and (3) poorer recognition of the visual training stimuli compared to all other groups.

In Study 2, participants listened to a non-native (Japanese) accent in the audiovisual or auditory modality to test whether a greater amount of *continuous* exposure to audiovisual stimuli (without separate training) would reveal a difference in adaptation between the two modalities. This design enabled us to examine the overall amount of adaptation, as well as adaptation at different time points in the experiment (for example, the presence of audiovisual speech cues may afford benefits to recognition of accented speech in earlier compared with later trials; Janse and Adank, 2012). In addition, participants' eye movements were recorded to verify that they were predominantly looking at the speaker's face. As in Study 1, if audiovisual cues *are* beneficial to perceptual adaptation to accented speech, we predicted that participants exposed to audiovisual accented speech would adapt to a greater extent than participants exposed to audio-only accented speech. Conversely, if audiovisual cues *are not* beneficial to perceptual adaptation to accented speech, we expected to observe no difference in perceptual adaptation for the audiovisual and auditory modalities in either study.

Study 1

Methods

Participants

One hundred and five students (26 male, *Median* = 20 years, age range 18–30 years) recruited from the University of Manchester, participated in the study. All participants were native British English speakers with no history of neurological, speech or language problems (self-declared), and gave their written informed consent. Participants were included if their corrected binocular vision was 6/6 or better using a reduced Snellens chart, and their stereoacuity was at least 60 s of arc using a TNO test. Participants' hearing was measured using pure-tone audiometry for the main audiometric frequencies in speech (0.5, 1, 2, and 4 kHz) in both ears. Any participant with a hearing threshold level greater than 20 dB for more than one frequency in either ear was excluded and did not participate in the study. We excluded one male participant based on the criteria for hearing, and four (one male, three female) based on the criteria for vision. We provided compensation of course credit or £7.50 for participation. The study was approved by The University of Manchester ethics committee.

Materials

We used 150 Institute of Electrical and Electronics Engineers (IEEE) Harvard sentences (IEEE, 1969) for our stimuli, and a 30-years-old male volunteer provided all recordings for the experiment. We transcribed and recorded 135 sentences in the novel accent, and randomly divided them into three lists, A, B or C. We recorded the remaining 15 sentences in the speaker's own British English accent to provide stimuli for a 'familiar accent' baseline test. We used a novel accent to avoid confounds from participant familiarity (that is, we could guarantee that none of our participants had ever encountered it before; see Adank et al., 2009), and to compare responses to the novel, unfamiliar accent with a familiar accent (our baseline measurement) from the same speaker (Adank and Janse, 2010). The novel accent (see Banks et al., 2015 for further details) was created by systematically modifying the vowel sounds of a Standard British English accent (**Table 1**). The accent was created using allophones from existing regional English accents (for example, Scottish or Irish) through an iterative process.

Training stimuli

Stimuli for the training sessions comprised six movies (three with and three without background noise), each comprising 45 video clips from one of the three novel-accented stimuli lists (A, B, and C). During recordings, the speaker looked directly at the camera with a neutral expression, and was asked to speak as naturally as possible. The recordings were made in a sound-treated laboratory with no natural light, using a High Definition Canon HV30 camera and Shure SM58 microphone. The camera was positioned ~1 m from the speaker to frame the head and shoulders, with a blue background behind the speaker. Video recordings were imported into iMovie 11, running on an Apple MacBook Pro, as

TABLE 1 | Phonetic description of the novel accent.

IPA	Example
ɪ → ɛ	sit → set
ɛ → ɪ	bet → bit
æ → ɛ	hat → het
ʌ → ʊ	cud → could
ɜː → ɛə	girl → gairl
aː → ɔː	dark → dork
ɒ → ɔː	hot → hawt
ɔː	door
uː	food
ʊ	good
ə	mother
iː	tree
ɛə → ɜː	hair → her
əʊ → aʊ	vote → vowt
aʊ → uː	how → hoo
ɛɪ → aɪ	way → wye
aɪ → ɔɪ	my → moy
ɪə	hear
ɔɪ	joy

large (960 × 540) digital video (.dv) files. Each recorded sentence was edited to create a 6-s video clip which were then compiled in a randomized order to create the training videos. Between each clip (sentence) there was a 7-s interval, during which the screen was black with a white question mark for 4 s (to indicate to participants they should respond) and a white fixation cross for 3 s (to indicate the next clip was imminent). Edited audio files (see Testing Stimuli, below) were re-attached to each video clip so that the normalized stereo tracks would be heard congruently with the video. For training conditions that included background noise, we added speech-shaped noise at a signal-to-noise ratio (SNR) of 0 dB to the audio files, using a custom script in Matlab software (R2010a, Mathworks, Inc.), before re-attaching them. Each movie was exported as a 960 × 540 MPEG-4 movie file with a bit-rate of 3269, in widescreen (16:9) ratio at 25 frames per second.

Testing stimuli

The audio track for each video clip (sentence) was extracted as an audio (.wav) file to be used for the auditory testing sessions. The experimenter checked all recordings and any that were not deemed suitable (for example due to mispronunciation or unnaturalness) were re-recorded in a second recording session. Audio files were normalized by equating the root mean square amplitude, resampled at 22 kHz in stereo, and cropped at the nearest zero crossings at voice onset and offset, using Praat software (Boersma and Weenink, 2012). The same procedure was used for the native-English recordings to produce stimuli for the familiar-accent baseline test.

We counterbalanced the presentation order of the novel-accented stimuli for the pre-training, training and post-training sessions across training groups; this was based on the sentence lists and followed the order ABC, CAB, and BCA. Each sentence was presented once per participant to avoid item-specific training effects. During the pre-training and post-training sessions, sentences were presented in a pseudo-random order per testing block and per participant, and the sentences used for the baseline and training sessions were presented in a fixed order.

Procedure

Figure 1 shows the experimental design in full. Participants first listened to the 15 familiar-accented (baseline) sentences to habituate them to the task and to the background noise. This was followed by the pre-training session, after which participants underwent training in one of five randomly assigned conditions ($N = 20$ per group): audiovisual, audio-only, visual (speech-reading), audiovisual + noise, audio-only + noise. Each participant was exposed to training stimuli from one of the three lists (A, B, or C) presented on a laptop computer. However, for the two audio-only groups the screen was not visible, and for the visual group, participants were asked to remove their headphones

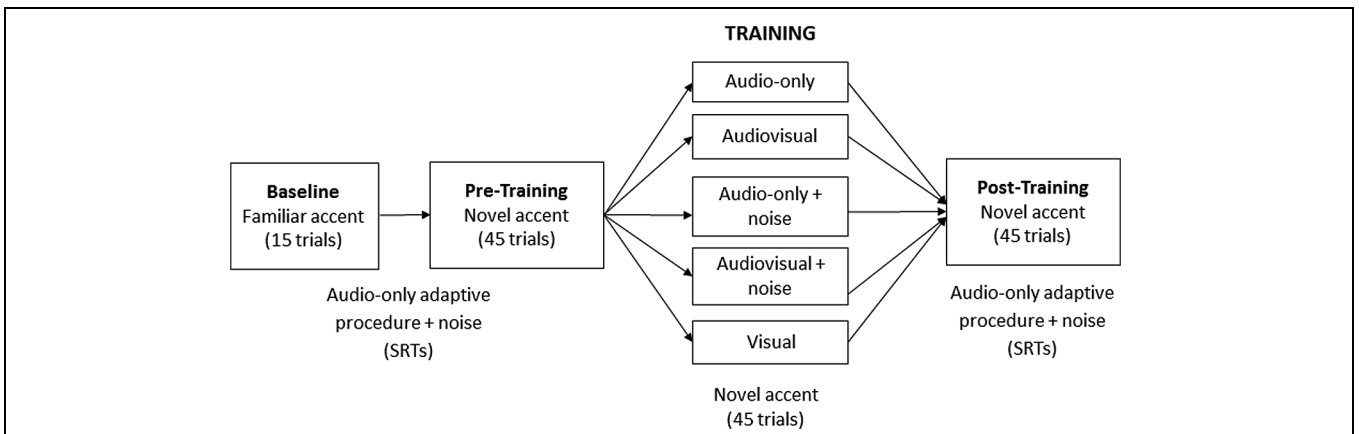


FIGURE 1 | Design for Study 1. The baseline session comprised 15 familiar-accented sentences; the pre-training, training and post-training session comprised 45 novel-accented sentences each.

and to speech-read each sentence. Each session (pre-training, training, and post-training) comprised 45 sentences.

Speech reception thresholds

For the baseline, pre-training and post-training sessions (but not for the training), we measured participants' recognition accuracy as speech reception thresholds (SRTs; Adank and Janse, 2010; Banks et al., 2015) in speech-shaped background noise, a sensitive measure which eliminates the need to equate starting accuracy between participants as it keeps recognition accuracy constant throughout. An adaptive staircase procedure (Plomp and Mimpen, 1979) varied the SNR per trial depending on the participants' response; that is, the SNR increased following an incorrect response, decreased following a correct response, or remained constant if a response was 50% correct. Thus, the SNR decreased as participants' performance improved (Baker and Rosen, 2001). The SNR varied in pre-determined steps of 8 dB for the first two changes and 2 dB thereafter, and maintained recognition accuracy (number of correctly repeated keywords) at 50%. The procedure was carried out using Matlab (R2010a). The mean SNR for all reversals indicated the SRT measurement for each participant, with an average of 21 reversals (SD = 5.4) per 45 trials.

Speech recognition task

Throughout the experiment, we instructed participants to repeat out loud as much of each sentence as they could in their normal voice and without imitating the accent. The experimenter scored participants' responses immediately after each trial, according to how many keywords (content or function words) they correctly repeated out of a maximum of four (for example, "a pot of tea helps to pass the evening"). Responses were scored as correct despite incorrect suffixes (such as -s, -ed, -ing) or verb endings; however, if only part of a word (including compound words) was repeated this was scored as incorrect (Dupoux and Green, 1997; Golomb et al., 2007; Banks et al., 2015). If a participant imitated the novel accent rather than responding in their own accent this was also scored as incorrect, as we could not make a clear judgment as to whether they had recognized the correct word.

All tests and training were carried out in a quiet laboratory in one session lasting ∼50 min. Auditory stimuli for the baseline and testing sessions were presented using Matlab software (R2010a, Mathworks, Inc.), and training stimuli were presented using iTunes 10.5.1 on an Apple MacBook Pro. Participants wore sound-attenuating headphones (Sennheiser HD 25-SP II) for the duration of the experiment, except during the visual (speech-reading) training. The experimenter adjusted the volume to a comfortable level for the first participant and then kept it at the same level for all participants thereafter.

Data Analysis

Perceptual adaptation was defined as the difference in SRTs before and after the training. We carried out a mixed-design ANOVA with a within-participant factor of testing session (two levels: pre- and post-training), and a between-group factor of training type (five levels: audio-only, audiovisual, visual-only, audio-only + noise, audiovisual + noise), was conducted on these difference scores. To investigate recognition of the novel accent in the different training modalities, we also analyzed accuracy scores (% correct keywords) from *within* the training session by conducting a one-way ANOVA (five levels: audio-only, audiovisual, visual-only, audio-only + noise, audiovisual + noise). To verify that baseline and pre-training measurements were equal across all groups, we carried out a one-way ANOVA for each data set with the between-group factor of training group (five levels). All *post hoc t*-tests carried out were two-tailed and we applied a Bonferroni correction for multiple comparisons. We identified two outliers in the data (one for the novel-accented SRTs and one for the baseline SRTs) with standardized residuals >3.29[1], and these scores were modified to the value of the group mean SRT plus two standard deviations. Unless otherwise stated, our data met all other assumptions for the parametric tests that we used.

Results

Table 2 shows the mean SRTs for the familiar-accented (baseline) speech, and mean pre- and post-training SRTs for the novel accent, per training group. As SRTs represent the SNR (dB) at which 50% recognition accuracy is achieved, higher levels reflect poorer performance. SRTs in all groups decreased following the training by ∼2 dB, indicating that participants' recognition of the accented speech improved over time and that perceptual adaptation took place. **Figure 2** shows the mean decrease in SRTs (amount of perceptual adaptation) following the training for each group. **Figures 3A–E** show a negative relationship between the amount of adaptation and pre-training SRTs; that is, participants who initially performed relatively worse adapted the most.

No significant differences were observed between groups for baseline SRTs (recognition of familiar-accented speech), or for pre-training SRTs (recognition of the novel-accented speech), confirming that the groups were equally matched for comparison. As expected, baseline SRTs across all five groups (M = 0.5 dB, SD = 1.68) were significantly lower than mean pre-training SRTs, across all groups (M = 7.7 dB, SD = 2.50), $t(99) = 29.19$, $p < 0.001$, confirming that the novel accent negatively affected participants' recognition in comparison to the

[1]In normally distributed data, z-scores would not be expected to be greater than 3.29.

TABLE 2 | Mean SRTs in dB per training group (Study 1).

	Familiar accent		Novel accent			
	Baseline SRT		Pre-training SRT		Post-training SRT	
Training group	*M*	SD	*M*	SD	*M*	SD
Audiovisual	0.4	1.68	7.6	2.33	5.0	2.66
Audio-only	0.2	1.42	7.9	2.36	5.0	1.88
Visual	0.6	1.75	7.8	2.56	6.0	1.89
Audiovisual + noise	0.9	2.19	7.4	3.13	5.9	3.46
Audio-only + noise	0.7	1.45	7.9	2.28	5.0	2.08
All groups	0.6	1.70	7.7	2.50	5.4	2.47

Training group N = 20.

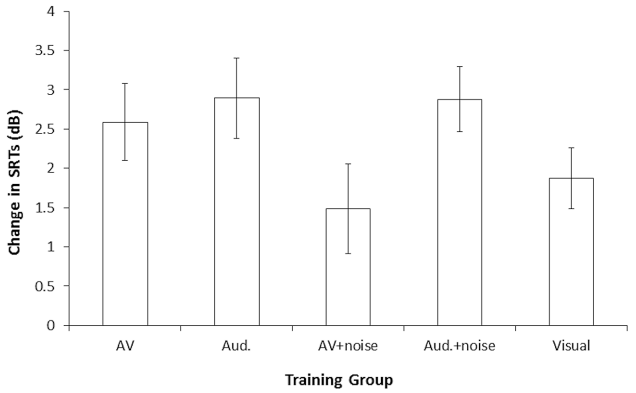


FIGURE 2 | Amount of perceptual adaptation to accented speech in Study 1: mean improvement in SRTs following training, per group (a higher change in SRTs indicates greater improvement). Error bars represent ± 1 SE. AV, audiovisual; Aud., audio-only.

familiar accent. We observed a main effect of testing session, $F(1,95) = 119.48$, $p < 0.001$, $\eta_p^2 = 0.56$. Paired-sample *t*-tests (Bonferroni correction, $p < 0.01$) confirmed that decreases in SRTs following the training were statistically significant in every group (see **Table 2**); thus, participants' recognition of the accented speech significantly improved between the two sessions. Neither the main effect of training group, nor the testing session × training type interaction, were significant ($ps > 0.05$).

A null finding may be interpreted in two ways: (1) that no effect is present in the population and the null hypothesis is true, or (2) that the data are inconclusive; however, significance testing cannot confirm these interpretations. Calculating Bayes factor (B) can, however, test whether the null hypothesis is likely, regardless of observed *p*-values. We calculated Bayes factor for differences in the amount of adaptation between all five groups (see **Figure 2** and **Table 3**). These analyses indicated that the null hypothesis (that there was no difference in adaptation between the groups) was supported for the following comparisons: audiovisual vs. audio, audiovisual vs. visual, audiovisual + noise vs. visual, audiovisual vs. audio + noise, and audio vs. audio + noise ($B < 0.33$; significant differences between these groups were predicted if our experimental hypotheses were true). All other comparisons indicated that data from this sample were inconclusive ($0.33 < B < 3.0$).

Analysis of the Training Data

To further investigate how the presence of audiovisual cues affected participants' recognition of the novel accent, we analyzed recognition accuracy in the five groups *during* the training (**Figure 4**). Analysis of these data revealed a significant effect of training condition, $F(4,95) = 331.47$, $p < 0.001$, $\eta_p^2 = 0.93$. Pairwise comparisons (Bonferroni correction, $p < 0.005$) confirmed that recognition accuracy was significantly lower in the visual group ($M = 1.4\%$, SD = 0.82) than in all other groups, $p < 0.001$. Recognition accuracy was also significantly higher in the audiovisual ($M = 85.2\%$, SD = 7.67) and audio-only ($M = 82.7\%$, SD = 7.17) groups compared to the audiovisual + noise ($M = 60.4\%$, SD = 11.90) and audio-only + noise ($M = 45.6\%$, SD = 10.01) groups, $ps < 0.001$. Recognition accuracy was significantly higher in the audiovisual + noise compared to the audio-only + noise group, $p < 0.001$. However, the marginal difference between the audiovisual and audio-only groups was not statistically significant, $p = 0.289$, and a Bayes factor calculation suggested that the data were inconclusive, $B = 0.30$ (uniform distribution, 0–30% limit).

Discussion

In Study 1, we investigated whether training with audiovisual or audio-only speech, with or without the presence of background noise, affected perceptual adaptation to a novel accent. As in previous studies of perceptual adaptation to accented speech (Clarke and Garrett, 2004; Bradlow and Bent, 2008; Maye et al., 2008; Adank and Janse, 2010; Gordon-Salant et al., 2010; Janse and Adank, 2012), we observed significant improvements in recognition of the novel accent over time, represented by a decrease in SRTs following the training.

Contrary to our predictions, there was no significant difference in the amount of adaptation between any of the groups; that is, the type of training had no effect on adaptation. Bayes factor suggested that non-significant differences in adaptation for four of the group comparisons (most importantly, audiovisual vs. audio-only) supported the null hypothesis. This would suggest that audiovisual cues do not benefit adaptation to accented speech better than audio-only or visual-only stimuli. However, for most of the group comparisons (particularly audio-only vs. visual), Bayes factor indicated that the data were inconclusive. We had included visual training as a control group, and predicted that training with audio-only stimuli would lead to greater adaptation in comparison – this would indicate that the training had been effective. However, the difference between these groups was inconclusive, and we therefore cannot ascertain whether the training was fully effective, or whether the lack of differences between groups was due to methodological reasons.

Analysis of data from the training session confirmed our predictions that recognition accuracy for the visual group would be considerably and significantly lower than all other groups, and that audiovisual cues would provide a benefit to recognition of the accented speech, as recognition accuracy was significantly higher in the audiovisual + noise group than in the audio-only + noise group. However, the same 'audiovisual benefit' was not present for participants carrying out training in quiet, although this null-effect was inconclusive – perhaps because accuracy was almost at ceiling level for these groups (Ross et al., 2007). Nevertheless, any effects observed during the training did not transfer to subsequent auditory testing, again suggesting that the training was not fully effective.

There are several possible explanations for this. Firstly, the timing of the training, and the length of the pre-training session, meant that participants had already begun adapting to the novel accent before the training. The training may therefore not have been fully beneficial at this stage. With longer exposure to the audiovisual stimuli at an earlier time point, we may have observed an effect of greater adaptation for this group. Secondly, inconsistency between the training and subsequent

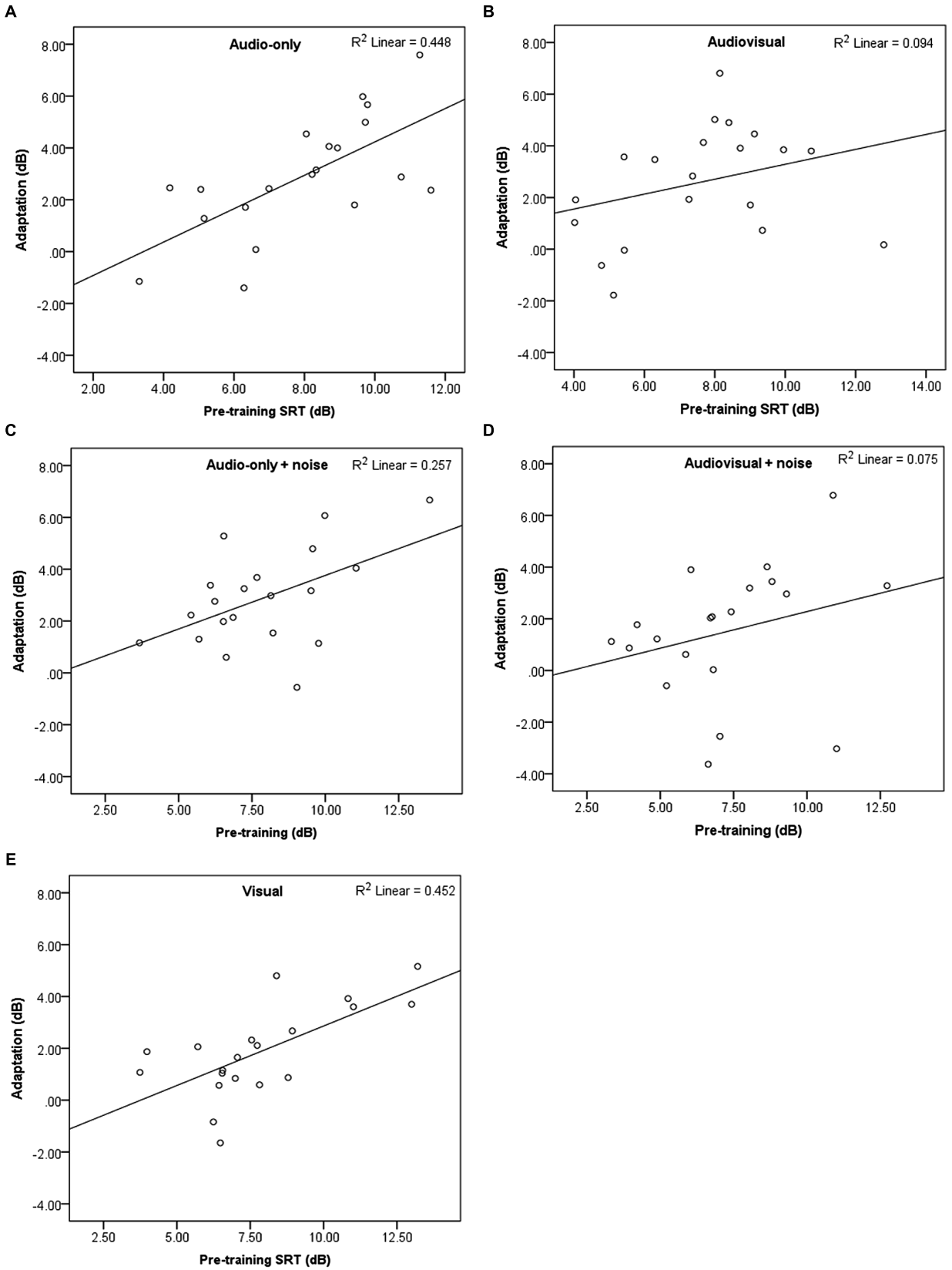


FIGURE 3 | (A–E) Scatterplots showing pre-training SRTs and amount of adaptation with linear fit, per group (Study 1).

TABLE 3 | Bayes factor (*B*) for comparisons of adaptation between groups in Study 1.

Group	AV	Audio	Visual	AV + noise
AV	–			
Audio	0.21*	–		
Visual	0.43	0.90	–	
AV + noise	0.85	1.68	0.24*	–
Audio + noise	0.20*	0.14*	1.10	1.97

*N = 20 per group. Calculations based on a uniform distribution with lower and upper limits of 0–6 dB change. B < 0.33 indicates results in favor of the null hypothesis; B > 3 indicates results in favor of the experimental hypothesis; intermediate values indicate that data is inconclusive. *B < 0.33.*

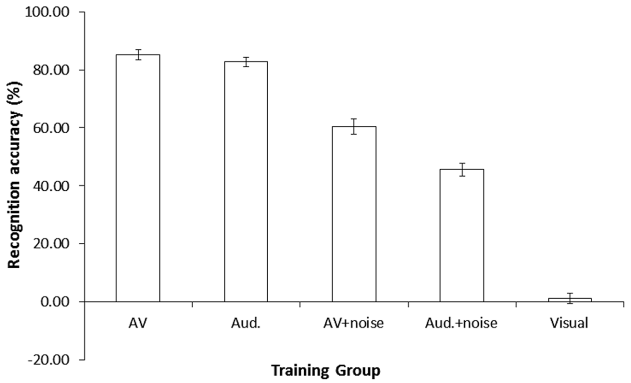


FIGURE 4 | Mean % recognition accuracy of accented speech per group, during the training session in Study 1. Error bars represent ± 1 SE. AV, audiovisual; Aud., audio-only.

testing sessions may have affected any benefits from the training, as consistency between training and subsequent testing can be beneficial to performance (Godden and Baddeley, 1975; Polyn et al., 2009). In fact, the switch to a separate training session may have been disruptive to adaptation. Thirdly, audiovisual cues from the particular speaker, or for the particular accent we used, may not have been sufficiently beneficial to improve perceptual adaptation. The relative benefit from audiovisual cues varies between different speakers (Kricos and Lesner, 1982, 1985), and this may also be the case for different accents. Indeed, Kawase et al. (2014) demonstrated that audiovisual cues vary in how much they can benefit recognition of non-native phonemes, in some cases even inhibiting recognition. Furthermore, Hazan et al. (2005) observed greater adaptation after audiovisual compared to audio-only training for *non-native* phonemes, whereas our novel accent was based on native (regional) English accents. We may therefore have observed a greater benefit to perceptual adaptation with audiovisual cues from a different speaker, and with a non-native accent.

To answer these remaining questions, we carried out a second study using a different experimental design, accent and speaker. In Study 2, we exposed participants to 90 sentences of unfamiliar accented speech in either the audiovisual or auditory modality without separate training, thus addressing concerns that the timing and length of the training, or inconsistency between training and testing sessions, affected the benefits gained from audiovisual cues in Study 1. Furthermore, this design allowed us to analyze the effects of audiovisual cues on adaptation at different stages of the experiment, for example during early compared with later trials, which may reveal more subtle effects (Janse and Adank, 2012). Secondly, we used a natural, non-native (Japanese) accent produced by a different speaker for our stimuli. Additionally, we recorded participants' eye movements using an eye-tracker to verify that they were continually looking at the speaker's face during testing. We increased the number of participants in each group to address any potential concerns that sample size prevented the effects in Study 1 from reaching statistical significance. By addressing these remaining questions, we hoped to clarify whether audiovisual speech cues can indeed benefit perceptual adaptation to unfamiliar accented speech.

Study 2

Methods

Participants

Sixty five young adults (five male, *Median* = 20.55 years, age range 18–30 years) recruited from the University of Manchester participated in the study, following the same procedure and exclusion criteria as Study 1. Two participants were excluded (one male, one female) due to data loss during the eye-tracking procedure (see Data Analysis for full details), and one female participant was excluded due to technical issues during the experiment.

Materials

Stimulus material consisted of 120 of the IEEE Harvard sentences (IEEE, 1969) that had been used in Study 1. A 30-year-old male native Japanese speaker recited 90 of them in a soundproofed laboratory, and these were recorded and edited using the same equipment and procedure as for Study 1. Speech-shaped background noise was added to the audio files using a custom Matlab script to create stimuli at SNRs of +4 to −4 dB. Background noise was included throughout to avoid ceiling effects associated with rapid perceptual adaptation to an unfamiliar accent (for example, Clarke and Garrett, 2004). For the audiovisual condition, the audio files were combined with the corresponding video clips using Experiment Builder software (SR Research, Mississauga, ON, Canada) to create congruous audiovisual stimuli. For the audio-only condition, a different static image of the speaker, taken from the video recordings, was displayed on screen simultaneously with each audio recording; this was to ensure that participants were processing auditory and visual information in both conditions. All stimuli were presented in a randomized order for each participant.

The native-accent baseline stimuli comprised the same 15 standard British English sentences from Study 1, plus an additional 15 recorded by the same speaker. We used 30 sentences to ensure that participants habituated to the background noise and task, as the SRT from this test would be used to set the SNR for presentation of the non-native accented stimuli. The baseline sentences were presented in a fixed order for all participants.

Procedure

All tests were carried out in a soundproofed booth in one session lasting ∼40 min. The familiar-accented baseline stimuli were presented and scored using Matlab software (R2010a, Mathworks, Inc.), through Sennheiser HD 25-SP II headphones, in the same adaptive staircase procedure used in Study 1 (see Speech Reception Thresholds for details). An Eyelink 1000 eye-tracker with Experiment Builder software (SR Research, Mississauga, ON, Canada) was used to present the accented stimuli and to record participants' eye movements. Participants wore the same headphones for the duration of the experiment, and sat with their chin on a chin rest facing the computer monitor. The experimenter adjusted the chin rest so that each participant's eyes were level with the top half of the display screen, which was positioned 30 cm from the chin rest. Eye movements were recorded by tracking the pupil and corneal reflection of the right eye at a sample rate of 1000 Hz. Calibration was carried out using a standard 9-point configuration before the start of the experiment, and 5 min after the start time. A drift-check was carried out immediately before each trial and calibration was performed again if required.

Participants were randomly allocated to either the audiovisual ($N = 32$) or audio-only ($N = 30$) condition. The experimenter set the volume for all stimuli at a comfortable level for the first participant, and kept it at the same level for all participants thereafter. Participants first listened to the 30 native-accented baseline sentences. The SRT acquired for this test was then used to set the SNR at which the accented stimuli were presented in the background noise, for each individual participant. The SRT was rounded to the nearest whole number (for example, if a participant's SRT for the familiar-accented speech was −1.3 dB, the SNR for the accented stimuli was set at −1 dB). This was intended to equate baseline recognition for the audiovisual group at ∼50% accuracy; however, we expected recognition to be lower for the audio-only group. This would allow us to verify the amount of 'benefit' provided by the audiovisual speech. In both conditions, participants were requested to watch the screen and to repeat each sentence following the same task and scoring procedure as in Study 1. Oral responses were recorded using a Panasonic lapel microphone attached to the chin rest, and responses were scored retrospectively by the experimenter. All 90 accented sentences were presented consecutively, and participants pressed the space bar to trigger each trial at their own pace.

Data Analysis

We measured recognition accuracy by calculating % correctly repeated keywords per sentence. To compare recognition accuracy between groups, and to analyze changes over time, we fitted a linear function to each participant's recognition data (Erb et al., 2012; Banks et al., 2015) using the equation $y = mx+b$, where y is the mean SRT, x is time (trial), m is the slope, and b is the intercept. The intercept of each participant's linear fit was used as the measurement of recognition accuracy, and the slope was used as the measurement of adaptation. We carried out t-tests and Bayes factor calculations to analyze effects of modality on recognition accuracy and perceptual adaptation. To confirm that participants in the audiovisual group were predominantly looking at the speaker's face, we created a semi-circular region of interest around this area, and calculated percent fixation time in this region for the duration of the stimulus presentation. We analyzed eye-tracking samples to check for data loss (for example due to blinks or head movements); trials with >20% data loss were excluded, and two participants who had >5 trials excluded were not included in our analyses (number of excluded trials: $M = 1.24$, SD = 3.06). For consistency, eye movement data were collected for both groups; however, as the data from the audio-only group is not relevant to this paper, these data will not be discussed further. All other analyses were conducted in the same way as in Study 1.

Results

Figure 5 shows mean recognition accuracy of the accented speech in the audiovisual and audio-only modalities, with linear fits. Recognition accuracy increased over time by a maximum of 10.8% (SD = 10.94) in the audiovisual group, and a maximum of 8.7% (SD = 13.61) in the audio-only group, suggesting that both groups adapted to the non-native accented speech. Recognition accuracy was consistently greater in the audiovisual group than the audio-only group, with a difference of ∼30% between the groups throughout the experiment. An independent-samples t-test confirmed that there was no significant difference in native-accented SRTs between the two groups, and that they were equally matched in their baseline ability to process non-native speech in background noise. **Figures 6A,B** show a negative relationship between the slope and intercept in each group indicating that, as in Study 1, participants with lower starting accuracy adapted the most.

There was a significant difference in the intercept for the audiovisual group ($M = 45.32$, SD = 9.52) and the audio-only group ($M = 14.44$, SD = 6.82); $t(57) = 13.82$, $p < 0.001$, $d' = 3.58$, confirming that recognition accuracy was significantly greater for the audiovisual group. However, there was no significant difference in slope between the audiovisual group ($M = 1.78$,

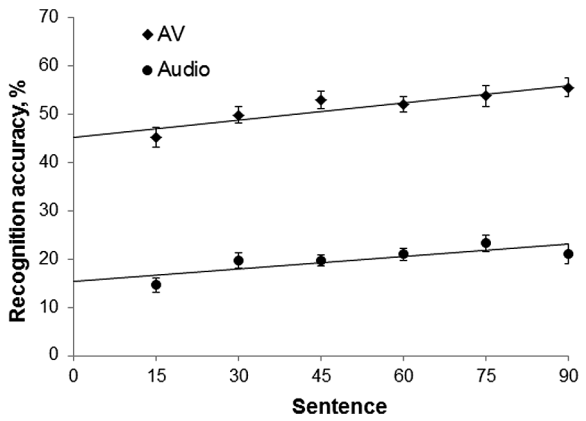


FIGURE 5 | Mean % recognition accuracy of accented speech in Study 2, per 15 sentences, per group, with linear fit. Error bars represent ± 1 SE. AV, audiovisual; Audio, audio-only.

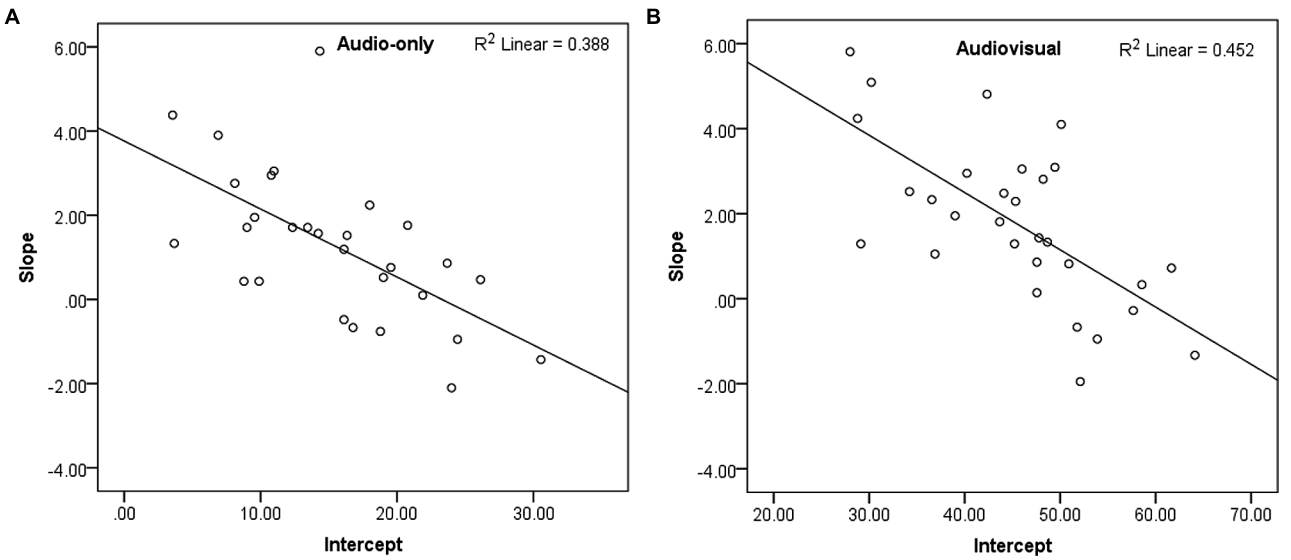


FIGURE 6 | (A,B) Scatterplots showing slope (adaptation) and intercept (baseline recognition accuracy) for recognition of the accented speech in Study 2, per group and with linear fit.

SD = 1.91) and the audio-only group (M = 1.27, SD = 1.77), $t(57) = 1.07$, $p = 0.291$, $d' = 0.28$. A Bayes factor calculation confirmed that the null hypothesis (that there was no difference in adaptation between the two groups) was likely, $B = 0.09$ (based on a uniform distribution and upper and lower limits of 0–20% improvement). Finally, analysis of the eye-tracking data confirmed that participants primarily looked at the speaker's face during presentation of the audiovisual stimuli (% gaze time on the speaker's face: $M = 100\%$, SD = 0.01%).

Discussion

Study 2 investigated whether perceptual adaptation to non-native accented speech differed when participants were exposed to audiovisual or audio-only stimuli. In comparison to Study 1, we exposed participants to the accented stimuli in either the audiovisual or audio-only modality without separate training. Participants were now exposed to twice as many audiovisual sentences as the training groups in Study 1, and could potentially benefit from the audiovisual cues at all stages of the experiment. Participants also performed the task in consistent conditions throughout the experiment without interruption, rather than in different modalities for testing and training. We used a Japanese accent and a different speaker for our stimuli to test whether audiovisual cues were more beneficial for recognizing a non-native accent (in comparison to the novel accent used in Study 1). Lastly, we recorded participants' eye gaze to confirm that they looked predominantly at the speaker's face.

As in Study 1, recognition accuracy of the accented speech significantly improved over time. We observed a maximum increase of ∼10%, which is similar to previous studies of perceptual adaptation to accented speech (Bradlow and Bent, 2008; Gordon-Salant et al., 2010; Janse and Adank, 2012). As predicted, participants exposed to audiovisual stimuli had better overall recognition of the foreign-accented speech in noise than those exposed to audio-only stimuli. This replicates previous findings that audiovisual speech cues can improve recognition of accented speech in noise (Janse and Adank, 2012; Yi et al., 2013). However, we found no significant difference in the amount of perceptual adaptation between the audiovisual and audio-only groups at any stage of the experiment. If audiovisual cues were beneficial to perceptual adaptation of accented speech (in comparison to audio-only cues), we expected to observe a statistically significant difference.

Overall Discussion

In the two studies described here, we investigated differences in perceptual adaptation to accented speech with audiovisual or audio-only stimuli. Study 1 employed an offline training design and a novel accent, while participants in Study 2 were exposed to a non-native accent in either modality without separate training. In both studies, we observed a benefit from audiovisual stimuli to recognition of the accented speech in noise. However, neither study demonstrated that audiovisual stimuli can improve perceptual adaptation to accented speech when compared to audio-only stimuli; furthermore, findings from Study 2 supported the null hypothesis.

Audiovisual Cues do not Improve Perceptual Adaptation to Accented Speech

We predicted that listeners would perceptually adapt to accented speech more when exposed to audiovisual stimuli, compared to just audio-only stimuli. We hypothesized that listeners would benefit from improved overall perception of the accented speech when visual cues were present (Arnold and Hill, 2001; Janse and Adank, 2012; Yi et al., 2013; Kawase et al., 2014), and would

therefore be better able to disambiguate the unfamiliar phonetic pattern of the accent, and map it to the correct lexical items more successfully.

In Study 1, there was no significant different in adaptation between any of the groups. Bayes calculations indicated that there was indeed no effect present between the audiovisual and audio-only groups, however, much of the data was inconclusive and the training may therefore have not been fully effective. We argued that this may have been due to: (1) the length or timing of the training, (2) inconsistencies between the training and testing sessions, or (3) the specific accent or speaker. Nevertheless, after addressing these concerns in the design of Study 2, there was still no clear advantage for perceptual adaptation to accented speech with audiovisual cues. In fact, Bayes analyses suggested that the data from Study 2 support the null hypothesis – that is, the presence of visual cues does not benefit adaptation to accented speech.

Our results support previous findings by Janse and Adank (2012), who observed no significant difference in adaptation between audiovisual and audio-only accented sentences in older adults. However, our results conflict with the findings of Hazan et al. (2005), who observed that audiovisual cues *can* improve perceptual adaptation to individual non-native phonemes. These conflicting results suggest that, although audiovisual cues may help listeners to perceptually learn individual speech sounds (as in Hazan et al., 2005), this benefit does not generalize to longer items of accented speech such as sentences (as used in the present study), perhaps reflecting the increased difficulty of speech-reading longer items (Grant and Seitz, 1998; Sommers et al., 2005).

Our results suggest that perceptual adaptation to accented speech is a robust ability that is not necessarily affected by the perceptual quality of the speech, as our participants adapted to the accented speech equally in conditions with or without visual cues that improved intelligibility. Indeed, Bradlow and Bent (2008) have demonstrated that the relative intelligibility of an accent (and therefore the perceived quality of the perceptual input) does not necessarily influence the amount that listeners can adapt to it. Perceptual adaptation to accented speech may therefore be primarily driven by factors internal to the listener rather than the perceptual environment, for example statistical learning (Neger et al., 2014) or cognitive abilities (Adank and Janse, 2010; Janse and Adank, 2012; Banks et al., 2015). However, it is possible that audiovisual cues benefit listeners in ways that we did not measure in the present studies, for example in terms of listening effort – that is, the presence of audiovisual cues may have reduced the effort associated with processing accented speech (Van Engen and Peelle, 2014). A more sensitive measure such as response times may have revealed a benefit from the audiovisual cues, although this was not the case for older adults (Janse and Adank, 2012).

Some limitations to the present findings should also be acknowledged. Firstly, a benefit from audiovisual cues may be present with more exposure. Indeed, a significant benefit from audiovisual cues has been observed for perceptual adaptation to noise-vocoded speech after exposure to a greater number of stimuli than in the present two studies (Pilling and Thomas, 2011). Secondly, the audio-only group in Study 2 had a lower baseline level of recognition accuracy than the audiovisual group (15% compared to 45% accuracy); this was intentional and allowed us to confirm that the presence of audiovisual speech cues from our speaker was beneficial to performance. However, it left more room for improvement in the audio-only group and potentially impacted the amount of adaptation our participants achieved, as in both groups poorer performers adapted the most (see **Figures 6A,B**). A comparison of adaptation to audiovisual and audio-only accented speech, with baseline recognition equated in both groups, may produce different results.

Audiovisual Cues Benefit Recognition of Accented Speech in Noise

Results from both studies replicate previous findings that audiovisual cues can benefit recognition of accented speech in noise when compared to only auditory cues (Arnold and Hill, 2001; Janse and Adank, 2012; Yi et al., 2013; Kawase et al., 2014). We observed a difference in recognition accuracy of ∼30% between the two groups in Study 2, and 15% between the two groups in Study 1 (during training with background noise). It is likely that visual cues from a speaker's facial movements help the listener to identify ambiguous or unclear phonemes by constraining the possible interpretations, or perhaps helping to identify prosodic cues (Swerts and Krahmer, 2008). Nevertheless, in both studies, we only observed greater recognition accuracy for the audiovisual groups when background noise was present, suggesting that benefits may have been related to compensation for the background noise, rather than the accented speech *per se*. Particularly, in Study 1 we did not observe a significant difference in recognition accuracy between the audiovisual and audio-only training groups when the stimuli were presented in quiet. However, recognition accuracy for these training groups was almost at ceiling level and the additional perceptual input from the audiovisual cues may therefore have been redundant, as the perceived clarity of the auditory signal can influence the benefits gained from audiovisual speech cues (Ross et al., 2007).

Listeners can perceptually adapt to accented speech very rapidly, even after exposure to a few sentences (cf. Clarke and Garrett, 2004), and this poses a practical limitation to studies of perceptual adaptation to, or recognition of, accented speech. As in the present studies, the most commonly used method to avoid ceiling effects is to add background noise, and this is the context in which an audiovisual benefit to accented sentences has previously been observed (Janse and Adank, 2012; Yi et al., 2013). However, two studies have also demonstrated this effect with audiovisual stimuli presented in quiet. Kawase et al. (2014) investigated adaptation to audiovisual accented phonemes in quiet; however, removing any lexical or semantic information increases the task difficulty, but perhaps does not reflect an ecologically valid context. Arnold and Hill (2001) used longer speech passages and a semantic comprehension task to assess the contribution of audiovisual cues; but, the task may have reflected semantic memory processes rather than speech recognition *per se*, and the result has not since been replicated. The extent

to which audiovisual cues can benefit recognition of accented speech in optimal, quiet listening conditions remains, therefore, to be confirmed.

Finally, we observed different amounts of audiovisual benefit between the two studies. This may be explained by differences in the speaker and accent used. Kawase et al. (2014) observed that audiovisual speech affects the perception of non-native phonemes to varying degrees; it is therefore likely that different accents result in varying benefits from visual speech cues. Furthermore, visemes (the visual equivalent of phonemes) from different speakers can vary in intelligibility (for example, Kricos and Lesner, 1982, 1985), possibly resulting in different benefits from our two speakers. Our results therefore add to existing evidence that being face-to-face with a speaker does not always benefit the listener to the same extent.

Conclusion

The present studies demonstrate that audiovisual speech cues do not benefit perceptual adaptation to accented speech – that is, observing audiovisual cues from a speaker's face does not lead to greater improvements in recognition of accented speech over time, when compared to listening to auditory speech alone. Audiovisual cues may still provide benefits to recognition of accented speech in noisy listening conditions, as we found a benefit to recognition of both types of accented speech in noise in comparison to audio-only speech. However, our results also demonstrate that the benefits obtained from audiovisual speech cues vary greatly, and the extent to which they benefit recognition of accented speech, as opposed to background noise, still needs to be clarified.

Acknowledgments

This work was funded by a Biotechnology and Biological Sciences Research Council research studentship awarded to BB, and by The University of Manchester. The authors thank Stuart Rosen for supplying the adaptive staircase program and the speech-shaped noise.

References

Adank, P., Evans, B. G., Stuart-Smith, J., and Scott, S. K. (2009). Comprehension of familiar and unfamiliar native accents under adverse listening conditions. *J. Exp. Psychol. Hum. Percept. Perform.* 35, 520–529. doi: 10.1037/a0013552

Adank, P., and Janse, E. (2010). Comprehension of a novel accent by young and older listeners. *Psychol. Aging* 25, 736–740. doi: 10.1037/A0020054

Arnold, P., and Hill, F. (2001). Bisensory augmentation: a speechreading advantage when speech is clearly audible and intact. *Br. J. Psychol.* 92, 339–355. doi: 10.1348/000712601162220

Baker, R., and Rosen, S. (2001). Evaluation of maximum-likelihood threshold estimation with tone-in-noise masking. *Br. J. Audiol.* 35, 43–52.

Banks, B., Gowen, E., Munro, K., and Adank, P. (2015). Cognitive predictors of perceptual adaptation to accented speech. *J. Acoust. Soc. Am.* 137, 2015–2024. doi: 10.1121/1.4916265

Bernstein, L. E., Auer, E. T. Jr., Eberhardt, S. P., and Jiang, J. (2013). Auditory perceptual learning for speech perception can be enhanced by audiovisual training. *Front. Neurosci.* 7:34. doi: 10.3389/fnins.2013.00034

Bertelson, P., Vroomen, J., and de Gelder, B. (2003). Visual recalibration of auditory speech identification: a McGurk aftereffect. *Psychol. Sci.* 14, 592–597. doi: 10.1046/j.0956-7976.2003.psci_1470.x

Boersma, P., and Weenink, D. (2012). *Praat: Doing Phonetics by Computer. Version 5.3.05*. Available at: www.praat.org (Last viewed 19/2/2012).

Bradlow, A. R., and Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition* 106, 707–729. doi: 10.1016/j.cognition.2007.04.005

Bruce, C., To, C.-T., and Newton, C. (2012). Accent on communication: the impact of regional and foreign accent on comprehension in adults with aphasia. *Disabil. Rehabil.* 34, 1024–1029. doi: 10.3109/09638288.2011.631680

Clarke, C. M., and Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *J. Acoust. Soc. Am.* 116, 3647–3658. doi: 10.1121/1.1815131

Cristia, A., Seidl, A., Vaughn, C., Schmale, R., Bradlow, A. R., and Floccia, C. (2012). Linguistic processing of accented speech across the lifespan. *Front. Psychol.* 3:479. doi: 10.3389/fpsyg.2012.00479

Dupoux, E., and Green, K. (1997). Perceptual adjustment to highly compressed speech: effects of talker and rate changes. *J. Exp. Psychol. Hum. Percept. Perfor.* 23, 914–927. doi: 10.1037//0096-1523.23.3.914

Erb, J., Henry, M. J., Eisner, F., and Obleser, J. (2012). Auditory skills and brain morphology predict individual differences in adaptation to degraded speech. *Neuropsychologia* 50, 2154–2164. doi: 10.1016/j.neuropsychologia.2012.05.013

Erber, N. P. (1975). Auditory-visual perception of speech. *J. Speech Hear. Disord.* 40, 481–492. doi: 10.1044/jshd.4004.481

Godden, D. R., and Baddeley, A. D. (1975). Context-dependent memory in 2 natural environments - land and underwater. *Br. J. Psychol.* 66, 325–331. doi: 10.1111/j.2044-8295.1975.tb01468.x

Golomb, J. D., Peelle, J. E., and Wingfield, A. (2007). Effects of stimulus variability and adult aging on adaptation to time-compressed speech. *J. Acoust. Soc. Am.* 121, 1701–1708. doi: 10.1121/1.2436635

Gordon-Salant, S., Yeni-Komshian, G. H., Fitzgibbons, P. J., and Schurman, J. (2010). Short-term adaptation to accented English by younger and older adults. *J. Acoust. Soc. Am.* 128, EL200–EL204. doi: 10.1121/1.3486199

Grant, K. W., and Seitz, P. F. (1998). Measures of auditory-visual integration in nonsense syllables and sentences. *J. Acoust. Soc. Am.* 104, 2438–2450. doi: 10.1121/1.423751

Grant, K. W., Walden, B. E., and Seitz, P. F. (1998). Auditory-visual speech recognition by hearing-impaired subjects: consonant recognition, sentence recognition, and auditory-visual integration. *J. Acoust. Soc. Am.* 103, 2677–2690. doi: 10.1121/1.422788

Hazan, V., Sennema, A., Faulkner, A., Ortega-Llebaria, M., Iba, M., and Chung, H. (2006). The use of visual cues in the perception of non-native consonant contrasts. *J. Acoust. Soc. Am.* 119, 1740–1751. doi: 10.1121/1.2166611

Hazan, V., Sennema, A., Iba, M., and Faulkner, A. (2005). Effect of audiovisual perceptual training on the perception and production of consonants by Japanese learners of English. *Speech Commun.* 47, 360–378. doi: 10.1016/j.specom.2005.04.007

IEEE. (1969). IEEE recommended practices for speech quality measurements. *IEEE Trans. Audio Electroacoustics* 17, 227–246.

Janse, E., and Adank, P. (2012). Predicting foreign-accent adaptation in older adults. *Q. J. Exp. Psychol.* 65, 1563–1585. doi: 10.1080/17470218.2012.658822

Jongman, A., Wang, Y., and Kim, B. H. (2003). Contributions of semantic and facial information to perception of nonsibilant fricatives. *J. Speech Lang. Hear. Res.* 46, 1367–1377. doi: 10.1044/1092-4388(2003/106)

Kawase, S., Hannah, B., and Wang, Y. (2014). The influence of visual speech information on the intelligibility of English consonants produced by non-native speakers. *J. Acoust. Soc. Am.* 136, 1352–1362. doi: 10.1121/1.4892770

Kawase, T., Sakamoto, S., Hori, Y., Maki, A., Suzuki, Y., and Kobayashi, T. (2009). Bimodal audio-visual training enhances auditory adaptation process. *Neuroreport* 20, 1231–1234. doi: 10.1097/WNR.0b013e32832fbef8

Kraljic, T., and Samuel, A. G. (2005). Perceptual learning for speech: is there a return to normal? *Cogn. Psychol.* 51, 141–178. doi: 10.1016/j.cogpsych.2005.05.001

Kraljic, T., and Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychon. Bull. Rev.* 13, 262–268. doi: 10.3758/bf03193841

Kricos, P. B., and Lesner, S. A. (1982). Differences in visual intelligibility across talkers. *Volta Rev.* 84, 219–225.

Kricos, P. B., and Lesner, S. A. (1985). Effect of talker differences on the speechreading of hearing-impaired teenagers. *Volta Rev.* 87, 5–14.

MacLeod, A., and Summerfield, Q. (1987). Quantifying the contribution of vision to speech perception in noise. *Br. J. Audiol.* 21, 131–141. doi: 10.3109/03005368709077786

Maye, J., Aslin, R. N., and Tanenhaus, M. K. (2008). The weckud wetch of the wast: lexical adaptation to a novel accent. *Cogn. Sci.* 32, 543–562. doi: 10.1080/03640210802035357

Munro, M. J., and Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of 2nd-Language learners. *Lang. Learn.* 45, 73–97. doi: 10.1111/j.1467-1770.1995.tb00963.x

Neger, T. M., Rietveld, T., and Janse, E. (2014). Relationship between perceptual learning in speech and statistical learning in younger and older adults. *Front. Hum. Neurosci.* 8:268. doi: 10.3389/fnhum.2014.00628

Norris, D., McQueen, J. M., and Cutler, A. (2003). Perceptual learning in speech. *Cogn. Psychol.* 47, 204–238. doi: 10.1016/S0010-0285(03)00006-9

Pilling, M., and Thomas, S. (2011). Audiovisual cues and perceptual learning of spectrally distorted speech. *Lang. Speech* 54, 487–497. doi: 10.1177/0023830911404958

Plomp, R., and Mimpen, A. M. (1979). Speech-reception threshold for sentences as a function of age and noise-level. *J. Acoust. Soc. Am.* 66, 1333–1342. doi: 10.1121/1.383554

Polyn, S. M., Norman, K. A., and Kahana, M. J. (2009). A context maintenance and retrieval model of organizational processes in free recall. *Psychol. Rev.* 116, 129–156. doi: 10.1037/a0014420

Reisberg, D., McLean, J., and Goldfield, A. (1987). "Easy to hear but hard to understand: a lip-reading advantage with intact auditory stimuli," in *Hearing by Eye: The Psychology of Lip-reading*, eds R. Campbell and B. Dodd (Hillsdale, NJ: Lawrence Erlbaum), 97–114.

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., and Foxe, J. J. (2007). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environment. *Cereb. Cortex* 17, 1147–1153. doi: 10.1093/cercor/bhl024

Sommers, M. S., Tye-Murray, N., and Spehar, B. (2005). Auditory-visual speech perception and auditory-visual enhancement in normal-hearing younger and older adults. *Ear Hear.* 26, 263–275. doi: 10.1097/00003446-200506000-00003

Stein, B. E., and Meredith, M. A. (1993). *The Merging of the Senses*. Cambridge, MA: MIT Press.

Sumby, W. H., and Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *J. Acoust. Soc. Am.* 26, 212–215. doi: 10.1121/1.1907309

Swerts, M., and Krahmer, E. (2008). Facial expression and prosodic prominence: effects of modality and facial area. *J. Phon.* 36, 219–238. doi: 10.1016/j.wocn.2007.05.001

Van Engen, K. J., and Peelle, J. E. (2014). Listening effort and accented speech. *Front. Hum. Neurosci.* 8:577. doi: 10.3389/fnhum.2014.00577

Vroomen, J., Keetels, M., van Linden, S., de Gelder, B., and Bertelson, P. (2004). Selective adaptation and recalibration of auditory speech by lipread information: dissipation. *Speech Commun.* 44, 55–61. doi: 10.1016/j.specom.2004.03.009

Wayne, R. V., and Johnsrude, I. S. (2012). The role of visual speech information in supporting perceptual learning of degraded speech. *J. Exp. Psychol. Appl.* 18, 419–435. doi: 10.1037/a0031042

Yi, H.-G., Phelps, J. E. B., Smiljanic, R., and Chandrasekaran, B. (2013). Reduced efficiency of audiovisual integration for nonnative speech. *J. Acoust. Soc. Am.* 134, EL387–EL393. doi: 10.1121/1.4822320

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

CHAPTER 5

EYE GAZE DURING RECOGNITION OF AUDIOVISUAL NOISE-VOCODED SPEECH

Manuscript to be submitted to the Journal of Experimental Psychology: Human Perception and Performance

Additional material relating to this chapter is presented in Appendix A.

Eye gaze during recognition of audiovisual noise-vocoded speech

Briony Banks, Emma Gowen, Kevin Munro and Patti Adank

University of Manchester

Author Note

Briony Banks and Kevin Munro, School of Psychological Sciences, University of Manchester; Emma Gowen, Faculty of Life Sciences, University of Manchester; Patti Adank, Speech, Hearing and Phonetic Sciences, University College London.

This research was supported by a BBSRC Doctoral Training Grant Studentship awarded to Briony Banks.

Correspondence regarding this article should be addressed to Briony Banks, Neuroscience and Aphasia Research Unit, School of Psychological Sciences, The University of Manchester, Zochonis Building, Brunswick Street, Manchester, M13 9LP. E-mail: briony.banks@manchester.ac.uk

Abstract

Listeners use visual speech cues to improve speech recognition in adverse listening conditions. However, it is not known exactly when listeners gain and use visual cues in these contexts, or whether this varies over time. We investigated when and for how long listeners directed their eye gaze towards a speaker's mouth, to determine when listeners gain and use visual speech cues during 1) recognition of individual noise-vocoded sentences, and 2) a longer period of perceptual adaptation to noise-vocoded speech. We additionally investigated whether measurements of eye gaze towards a speaker's mouth are related to successful recognition of noise-vocoded speech.

Fifty-eight participants perceived either audiovisual or audio-only noise-vocoded sentences in a speech recognition task. The audio-only group also viewed static images of the speaker's face to allow for direct comparison of eye gaze between the two groups. Data from the audiovisual group revealed three clear patterns: 1) the percentage and duration of fixations on the mouth increased during recognition of individual sentences; 2) as participants adapted to the noise-vocoded speech, the duration of fixations on the speaker's mouth decreased; 3) longer fixations on the mouth were associated with better recognition of the noise-vocoded speech. Conversely, the audio-only group consistently looked at the speaker's eyes more than the mouth, and eye gaze was not related to recognition accuracy; however, fixations also increased in duration during recognition of individual sentences in this group. Results confirm that eye gaze (specifically, longer fixations on a speaker's mouth) is related to successful recognition of unfamiliar audiovisual speech. Changes in the percentage and duration of fixations on the mouth suggest that listeners' use of visual speech cues varies over time, according to their needs; however, these changes could also reflect variation in cognitive effort.

Keywords: speech recognition, perceptual adaptation, audiovisual speech, eye movements, eye-tracking.

Eye gaze during recognition of audiovisual noise-vocoded speech

Since the seminal study by Sumby & Pollack (1954), the benefits of audiovisual speech perception have been extensively researched. Multiple studies have reported better recognition of speech in adverse listening conditions when a listener can see the speaker's face, compared to auditory processing alone (Erber, 1975; Grant et al., 1998; A. Macleod & Summerfield, 1987; Sommers et al., 2005). In adverse conditions, listeners primarily benefit from visual cues from a speaker's mouth articulations, for example to distinguish the place of articulation for different consonants (Summerfield, 1987). The benefits from audiovisual speech have predominantly been observed for speech in background noise, but also for recognition of noise-vocoded speech – an acoustic distortion which alters the spectral information of speech, making it sound like a 'noisy whisper' and thus more difficult to recognise (Shannon et al., 1995). Studies have shown that audiovisual cues not only benefit recognition of noise-vocoded speech, but also that they can lead to greater perceptual adaptation – that is, listeners' recognition has improved more with the presence of audiovisual cues than with auditory cues alone (Bernstein et al., 2013; T. Kawase et al., 2009; Pilling & Thomas, 2011). However, we do not know exactly when listeners gain and use visual cues during audiovisual speech recognition in a particular adverse listening condition, or during perceptual adaptation to unfamiliar speech; for example, listeners may rely more on visual cues when they initially encounter unfamiliar speech, to make predictions about the upcoming speech in a particular sentence, or to adapt to the unfamiliar speech type.

Examining where, and when, eye movements occur during recognition of a visual scene can reveal the cognitive processes underlying visual perception (for reviews, see Liversedge & Findlay, 2000; Rayner, 1998), for example indicating the salience of particular visual objects. Analysing listeners' eye gaze towards a speaker's mouth could thus potentially reveal when visual speech cues are most salient to the listener. Eye-tracking studies have shown that in optimal listening conditions, listeners tend to look primarily towards a speaker's eyes (Buchan et al., 2007, 2008; Vatikiotis-Bateson et al., 1998), most likely to gain social cues (for reviews, see Birmingham & Kingstone, 2009; Langton et al., 2000). However, in adverse conditions such as in the presence of background noise, listeners look increasingly towards the speaker's mouth (Buchan et al., 2007; 2008; Lansing & McConkie, 2003), and more so as levels of

background noise increase (Vatikiotis-Bateson et al., 1998). Lansing & McConkie (2003) further demonstrated that listeners shift their gaze towards a speaker's mouth at speech onset, and back to the eyes at speech offset, when perceiving clear but quiet audiovisual speech. Taken together, this evidence suggests that listeners specifically direct their eye gaze towards a speaker's mouth to make use of visual speech cues, and that eye gaze varies depending on the needs of the listener (for example, they look more towards the mouth when they need the input of visual speech cues due to an unclear auditory signal).

These observations fit well within the cognitive relevance framework of visual perception, which stipulates that the weight allocated to a particular visual feature is dependent on the information-gathering needs of the perceiver (Henderson, Malcolm, & Schandl, 2009). Indeed, Vo, Smith, Mital & Henderson (2012) observed that perceivers allocate their eye gaze to different areas of a speaker's face depending on the current task. From this framework, we can predict that eye gaze will vary even within the same listening context, as listeners will use visual speech cues when they are most beneficial to them. For example, during recognition of individual sentences, listeners may look more towards the mouth at the start of the sentence to 'tune in' to the unfamiliar speech or to make accurate predictions, but then less so as the sentence progresses. Equally, eye gaze patterns may change over a longer period of unfamiliar audiovisual speech recognition, as listeners adapt to the adverse listening condition. When perceivers are exposed to audiovisual noise-vocoded speech, for example, their recognition improves over time (Kawase et al., 2009; Pilling & Thomas, 2011); this process of perceptual adaptation may also correspond to a shift in eye gaze away from the mouth towards the eyes, as perceivers rely less on the visual speech cues as their performance improves. Eye gaze may therefore be a dynamic behaviour that varies not just according to external changes in the environment, but also according to internal changes within the listener such as perceptual learning. However, to our knowledge no studies of audiovisual speech perception have investigated patterns of eye gaze over a period of time – either during recognition of individual speech items such as sentences, or during perceptual adaptation to multiple speech items over an extended period of time.

Evidence from studies of audiovisual speech recognition in noise (Buchan et al., 2007; 2008; Lansing & McConkie, 2003; Vatikiotis-Bateson et al., 1998) further

suggest that eye gaze is related to speech recognition. Indeed, if perceivers look more at the mouth with increasing levels of background noise (Vatikiotis-Bateson et al., 1998), this would suggest that looking at the mouth is associated with poorer performance – that is, perceivers likely look at the speaker's mouth to compensate for their poor recognition. However, no such relationship has been observed. Several studies have examined correlations between eye gaze (for example, the number of fixations on the speaker's mouth) and speech recognition (Buchan et al., 2007; Everdell et al., 2007), as well as recognition of particular phonemes (Vatikiotis-Bateson et al., 1998), and speech-reading (Lansing & McConkie, 2003). Nevertheless, none of them reported significant correlations, even when controlling for speech-reading proficiency and item difficulty. In the first three of these studies, mean recognition accuracy for the audiovisual speech was almost at ceiling level (≥86%), while in the case of speech-reading (Lansing & McConkie, 2003), it was relatively low (20-30%). Furthermore, a different measure of eye gaze was analysed in the correlations reported in each of the studies cited here – percentage fixations on the mouth (Buchan et al., 2007), duration of fixations on the mouth (Everdell et al., 2007), percentage fixation time on the mouth (Lansing & McConkie, 2003), and spatial location of eye gaze (Vatikiotis-Bateson et al., 1998). It is thus unclear which measurement would be most likely to predict recognition accuracy. Using a relatively difficult speech recognition task to produce greater variance in recognition accuracy, and a variety of eye gaze measures, may therefore reveal a correlation.

The present study therefore addressed two questions in relation to eye gaze during recognition of unfamiliar audiovisual speech:

1) When do listeners gain and use visual speech cues – specifically, does eye gaze towards a speaker's mouth vary during recognition of individual sentences, and over a longer period of perceptual adaptation? We hypothesised that eye gaze would rapidly shift from the eyes towards the mouth during recognition of individual sentences, and that perceivers would look more towards the mouth at the start of sentences than at the end, in order to predict and adapt to the unfamiliar speech. Over the course of the experiment, we hypothesised that the amount of eye gaze towards the speaker's mouth would decrease as perceivers adapted to hearing the noise-vocoded speech, and therefore relied less on the visual speech cues.

2) Is eye gaze towards a speaker's mouth related to speech recognition? We predicted that eye gaze towards the speaker's mouth would be negatively correlated with their performance in the speech recognition task – that is, better performers would look less at the mouth, particularly in later trials following adaptation to the unfamiliar speech. Finally, we predicted that these patterns of eye gaze would only be observed for the audiovisual group, and not for the audio-only group, indicating that eye gaze towards the speaker's mouth reflected the use of visual speech cues.

We addressed these questions by comparing patterns of eye gaze in two groups of young, healthy participants, during recognition of noise-vocoded sentences. We exposed one group of participants to audiovisual sentences (that is, they listened to and watched a video of the speaker), while a control group listened to audio-only sentences while watching still images of the speaker's face; this was intended to control for the type and presence of visual information. We included a control group to infer whether patterns of eye gaze in the audiovisual group were specifically related to processing the audiovisual speech.

We used noise-vocoded speech for our stimuli (Shannon et al., 1995). Perceptual adaptation to noise-vocoded speech (that is, improved recognition over time), has been reliably demonstrated in the literature (Davis et al., 2005; Hervais-Adelman et al., 2008), and studies have shown that perceivers adapt more with audiovisual compared to audio-only speech cues (Bernstein et al., 2013; T. Kawase et al., 2009; Pilling & Thomas, 2011). We therefore also expected to observe greater adaptation in our audiovisual group compared to the audio-only group.

Methods

Participants

69 young adults (10 male, *Mdn* = 23 years, age range 19-30 years), recruited from the University of Manchester, participated in the study. All participants were native British English speakers with no history of neurological, speech or language problems (self-declared), and gave their written informed consent. Participants were included if their corrected binocular vision was 6/6 or better using a reduced Snellen

chart, and their stereoacuity was at least 60 seconds of arc using a TNO test. Participants' hearing was measured using pure-tone audiometry for the main audiometric frequencies in speech (0.5, 1, 2 and 4 kHz) in both ears. Any participant with a hearing threshold level greater than 20 dB for more than one frequency in either ear was excluded and did not participate in the study. Eleven participants in total (one male) were excluded; two based on the hearing criteria, two based on the visual criteria, five due to data loss during the eye tracking procedure (see Data Analysis for full details), one due to poor calibration during eye tracking, and two due to technical failure. We provided compensation of course credit or £7.50 for participation. The study was approved by The University of Manchester ethics committee.

Materials

Our stimuli consisted of 91 randomly selected Institute of Electrical and Electronics Engineers Harvard sentences (IEEE, 1969). Recordings were carried out in a sound-proofed laboratory using a Shure SM58 microphone and a High Definition Canon HV30 camera. A 26-year-old female native British English speaker recited the sentences, and was asked to look directly at the camera, to remain still and to maintain a neutral facial expression throughout the recordings, to minimise head movement (see Figure 1). Video recordings were imported into iMovie 11, running on an Apple MacBook Pro, as large (960 x 540) high definition digital video (.dv) files. Recordings were edited to create individual video clips for each sentence. These were checked by the experimenter and any that were not deemed suitable (for example due to mispronunciation) were re-recorded. The audio tracks for each clip were extracted as audio (.wav) files, then normalised by equating the root mean square amplitude, resampled at 22 kHz in stereo, cropped at the nearest zero crossings at voice onset and offset, and vocoded using Praat software (Boersma & Weenink, 2012) and custom scripts. Speech recordings were noise-vocoded according to Shannon et al (1995) using 4 frequency bands (cut-offs: 50 Hz → 369 Hz → 1160 Hz → 3124 Hz → 8000 Hz), selected to represent equal spacing along the basilar membrane (Greenwood, 1990). To create the still images to be displayed along with the audio files (for the audio-only group), screen shots (saved as TIFF files) were taken from the videos of the speaker in a variety of mouth positions, to make it evident that she was speaking. The still images, video files and the noise-vocoded audio files were then imported separately into

Experiment Builder software (SR Research, Ontario, Canada). In the audio-only condition, the still images of the speaker were displayed for the length of each audio file, and for the audiovisual condition the audio and video files were played congruously.

Procedure

We carried out the experiment in a sound-proofed testing booth in one session lasting approximately 40 minutes. Participants were randomly allocated into either the audiovisual or audio-only group. In both conditions, participants sat facing the screen approximately 50 cm from the monitor, with their chin on a chin-rest. They were asked not to move their head during the experiment and to look continuously at the screen. Before starting the experiment, the eye-tracker was calibrated for each participant (see 'Eye-tracking' for details). Participants first listened to one practice sentence (a clear version and a noise-vocoded version) that was not included in the experiment, to prepare them for hearing the unusual distortion. They then completed 90 trials with the remaining noise-vocoded sentences. Participants triggered the start of the experiment and each subsequent trial by pressing the space bar on the keyboard (that is, the experiment was self-paced). All stimuli were presented through Sennheiser HD 25-SP II headphones. The experimenter set the volume for all stimuli at a comfortable level for the first participant, and kept it at the same level for all participants thereafter. A Panasonic lapel microphone attached to the chin-rest recorded their verbal responses.

Speech recognition task. To measure speech recognition, we asked participants to repeat out loud as much of each sentence as they could. The experimenter retrospectively scored participants' responses according to how many keywords (content or function words) they correctly repeated out of a maximum of four. Responses were scored as correct despite incorrect suffixes (such as -s, -ed, -ing) or verb endings; however if only part of a word (including compound words) was repeated this was scored as incorrect (Dupoux & Green, 1997; Golomb et al., 2007).

Eye tracking. We used a desktop-mounted Eyelink 1000 eye-tracker with Experiment Builder software (SR Research, Ontario, Canada) to present all stimuli, and to record participants' eye movements. The pupil and corneal reflection of each participant's right eye were tracked at a sample rate of 1000 Hz, with a spatial

resolution of 0.01° RMS and average accuracy of 0.25°–0.5°. Calibration was carried out before the experiment using a standard 9-point configuration, and again 5 minutes after the experiment began. Each calibration was validated for accuracy, and was accepted if the average error was <1° and the maximum error was <1.5°. A drift check preceded each trial, and if the error between the computed fixation position and the on-screen target was >1.5°, calibration was again carried out to correct this drift.

Data Analysis

Measurements of eye gaze. Eye fixations (that is, any period of time when eye gaze is relatively still) reflect the perceiver's foveal field of vision and thus the area of greatest visual acuity. The number and duration of fixations can indicate where and to what extent a perceiver's visual attention is primarily directed at any given time (e.g. Christianson, Loftus, Hoffman, & Loftus, 1991). Fixation duration is also influenced by both lower and higher level cognitive processes, and has been associated with increased or effortful cognitive processing, for example during reading (Rayner, 1998). In particular, longer fixations on a speaker's mouth may indicate processing of visual speech cues (Buchan et al., 2007; 2008; Lansing & McConkie, 2003). We therefore selected three variables based on participants' fixations with which to analyse patterns of eye gaze: percent fixation time, percent fixations and fixation duration. Fixations were defined as any period that was not a saccade (saccades were defined as eye movements with velocity >30°/sec, acceleration >8000°/sec^2, and motion >0.1°). Percent fixation time was calculated as the summed duration of all fixations on an interest area divided by the total trial time. We selected this variable to compare the overall amount of time spent fixating on the eyes and mouth during perceptual adaptation, and thus which areas were of most interest to participants at particular time points. Percent fixations comprised the percentage of all fixations in a trial falling in the current interest area, while fixation duration was calculated as the mean duration of fixations in milliseconds. We selected these variables to indicate where listeners were looking at particular time points, and the degree of eye movement – that is, whether participants were fixating steadily (with fewer, longer fixations) or 'scanning' the visual scene (with more, shorter fixations). Measurements of eye gaze were computed using Data Viewer (SR Research, Ontario, Canada), and we calculated the mean of each variable per testing block, and per interest area.

Interest areas. For each video clip, we created two elliptical interest areas (IAs; see Figure 1). These comprised the eye area (extending from just below the speaker's eyebrows to the tip of the nose) and the mouth area (from the septum to just below the bottom lip). Eye gaze was then analysed based on these IAs to compare patterns of eye gaze between the two areas. We also created a third interest area which surrounded the speaker's face. This was only used to verify the proportion of eye gaze directed to the speaker's face rather than peripheral areas of the screen, and was not included in any other analyses.

Statistical analyses. We separated our data analysis into three parts each linked to one of our aims, and these are described in the following sections.

Part A. Is perceptual adaptation to noise-vocoded speech greater with audiovisual compared to audio-only cues?

To analyse recognition of the noise-vocoded speech, we divided all consecutive trials into 6 blocks of 15 trials each, and calculated mean percentage accuracy per testing block based on the number of correctly repeated keywords. We analysed differences between groups and changes in recognition accuracy over time (i.e. perceptual adaptation) by carrying out a mixed ANOVA with the between-group factor of group (2 levels: audiovisual, audio-only) and the within-participant factor of testing block (6 levels: blocks 1 - 6). Two-tailed paired-sample *t*-tests, with appropriate Bonferroni corrections, were conducted to compare differences in recognition accuracy between testing blocks.

Part B. When do perceivers of audiovisual noise-vocoded speech use visual speech cues?

Eye gaze during recognition of individual sentences. Seven temporal interest periods were created for the video/audio clips, each 500 ms in length (0-3500 ms from the start of the video or audio clip). A mixed-design ANOVA was then carried out for percent fixations and fixation duration during recognition of the noise-vocoded sentences, with the between-group factor of group (2 levels: audiovisual, audio-only) and the within-participant factors of IA (2 levels: eyes, mouth) and time (7 levels: 0-500 ms, 500-1000 ms, 1000-1500 ms, 1500-2000 ms, 2000-2500 ms, 2500-3000 ms, 3000-

3500 ms). Two-tailed paired-sample t-tests were also conducted to analyse differences between IAs and time-points/testing blocks, and two-tailed independent-samples t-tests to analyse differences between groups. Appropriate Bonferroni corrections were applied to these analyses.

Eye gaze during perceptual adaptation to noise-vocoded speech. To analyse eye gaze as participants adapted to the noise-vocoded speech (that is, over the course of the experiment), we divided all consecutive trials into 6 blocks of 15 trials each, and calculated mean percent gaze time, percent fixations, and fixation duration per testing block. We then carried out a mixed-design ANOVA for each eye gaze variable with the between-group factor of group (2 levels: audiovisual, audio-only) and the within-participant factors of IA (2 levels: eyes, mouth) and testing block (6 levels: blocks 1-6).

For all ANOVAs, if the assumption of homogeneity of variance was violated, degrees of freedom were corrected using a Greenhouse-Geisser correction. We also carried out two-tailed paired-sample t-tests to analyse specific differences between IAs and time-points/testing blocks, and two-tailed independent-samples t-tests to analyse differences between groups. Appropriate Bonferroni corrections were applied to these analyses.

Some data distributions for the eye gaze variables were significantly skewed, and could not be corrected through data transformation. In such cases, non-parametric statistics are usually recommended. However, we were specifically interested in data interactions (for example, between groups and IAs to infer whether patterns of eye gaze were specifically related to processing of visual speech cues), and there is no equivalent non-parametric test to specifically assess interactions. Parametric analyses can perform well even on skewed data, although false positives may be a risk (Glass, Peckham, & Sanders, 1972; Harwell, Rubinstein, Hayes, & Olds, 1992; Lix, Keselman, & Keselman, 1996). We therefore used parametric statistics for our analyses, but carried out additional non-parametric tests (included in Appendix A) to ensure that we were not reporting false positives. All non-parametric analyses revealed the same results as the parametric tests, apart from two analyses which reported significant effects not present in the parametric equivalent (see Results, Part B and Part C). Apart from these

exceptions, we can assume that the parametric analyses reported here accurately reflect effects within the data.

Part C. Is eye gaze related to recognition of audiovisual noise-vocoded speech?

Correlational analyses. We expected performance and patterns of eye gaze to change over time. We therefore analysed early and later trials separately to investigate the relationship between eye gaze and speech recognition. We calculated mean recognition accuracy, percent fixation time, percent fixations and fixation duration on the mouth between testing blocks 1-3 (early), and blocks 4-6 (later), for each group. We carried out Pearson's correlations between the variables of interest, and compared them between groups using Fisher's *z*.

Fixation duration over time: good and poorer performers. To analyse fixation duration for good and poorer performers in the audiovisual group, an additional analysis was carried out with the extra between-group factor of performance in the speech recognition task. Performance was classified as 'good' if mean recognition accuracy in the first testing block[1] was ≥ the median value (39%), and 'poor' if recognition accuracy was below this value. We carried out a mixed ANOVA for recognition accuracy and for fixation duration with the between-group factor of performance (2 levels: good and poor) and the within-participant factor of testing block (6 levels: blocks 1-6).

Results

Part A. Is perceptual adaptation to noise-vocoded speech greater with audiovisual compared to audio-only cues?

Figure 2 shows mean recognition accuracy for the noise-vocoded speech per testing block, for each group. Recognition was overall significantly better in the audiovisual group (M = 54%, SD = 2.0%) compared to the audio-only group (M = 35%, SD = 1.6%), $F(1, 57) = 57.24$, $p < 0.001$, $\eta_p^2 = 0.95$, and this difference was significant

[1] *NB* – the positive correlation between early and later testing blocks (see Tables 1a and 1b) indicated that good performers performed continuously well throughout the experiment, and vice versa.

in every testing block (*p*s < 0.008). Recognition accuracy significantly increased in both groups by approximately 19%, confirmed by a significant effect of testing block, $F(5, 285) = 34.53$, $p < 0.001$, $\eta_p^2 = 0.38$. In both groups, there were significant differences between block 1 and blocks 2, 3, 4, 5 and 6; between blocks 2 and 6, and blocks 3 and 6, *p*s < 0.003. A significant difference was also observed between blocks 2 and 5 in the audiovisual group, $p < 0.003$. In both groups, the greatest amount of improvement between any two consecutive blocks was between blocks 1 and 2 (audiovisual: $M = 9\%$, SD = 8.9%; audio-only: $M = 11\%$, SD = 10.4%).

We predicted that adaptation would be greater in the audiovisual group compared to the audio-only group; however, the group x testing block interaction was not significant ($p > 0.05$). We calculated Bayes factor (*B*) to establish whether this was due to sampling error or a genuine effect in the population. For a between-group comparison of the mean amount of adaptation between blocks 1 and 6 ($t(57) = 0.75$, $p = 0.382$), with a uniform distribution and an estimated effect of between 0-14% (based on data from Pilling & Thomas, 2011), $B = 0.34$, indicating that the data were inconclusive; that is, they could not substantially support the null or the experimental hypothesis (Dienes, 2014).

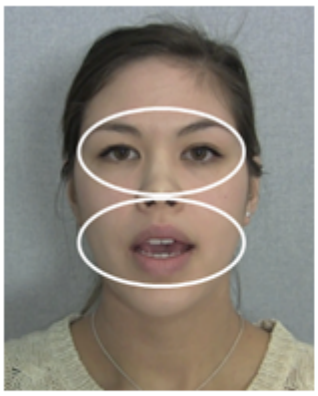


Figure 1. Image of the speaker with example interest areas.

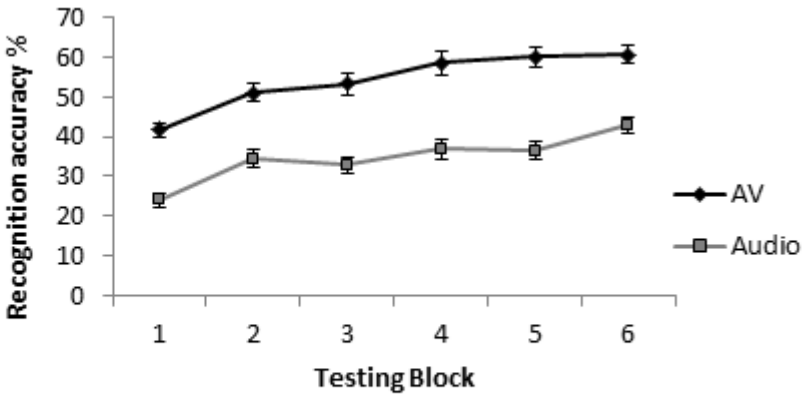


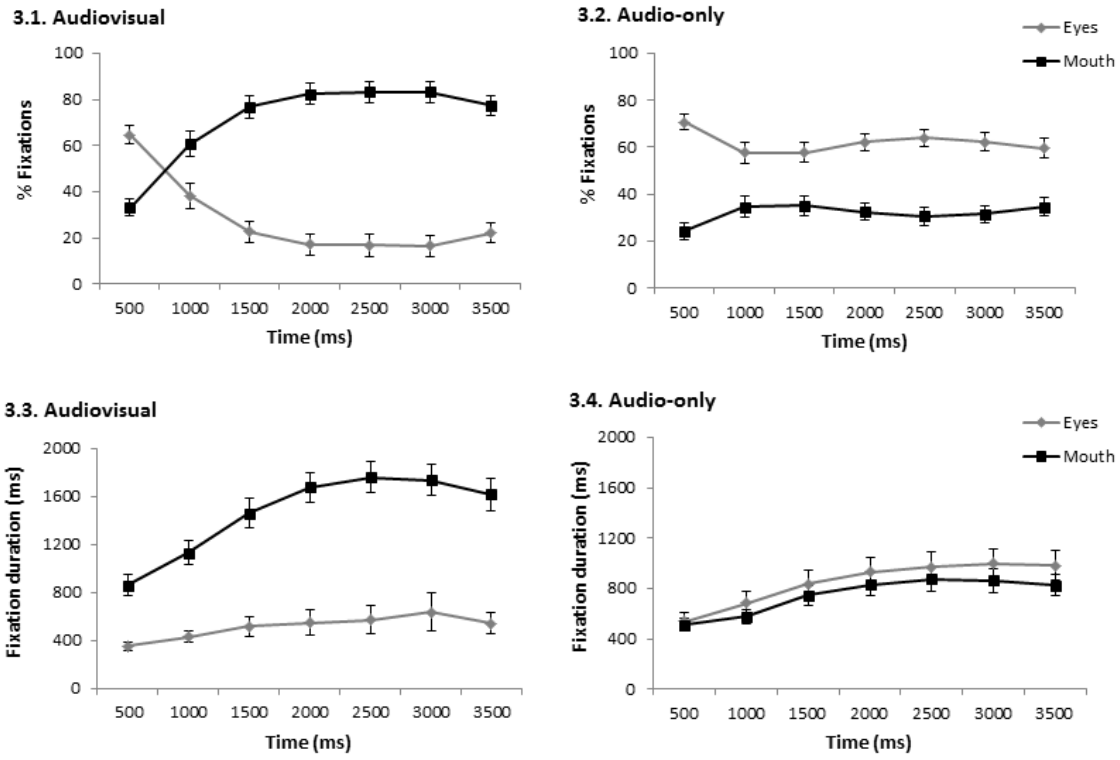
Figure 2. Mean recognition accuracy per testing block, per group. Error bars show ±1SE.

Part B. When do perceivers of audiovisual noise-vocoded speech use visual speech cues?

Eye gaze during recognition of individual sentences.

General patterns. Figures 3.1 – 3.4 show patterns of eye gaze (percent fixations and fixation duration) during presentation of the noise-vocoded sentences, for each group. Overall, the audiovisual group looked more at the speaker's mouth than the eyes, while the audio-only group looked more at the eyes. Changes over time were evident for percent fixations and fixation duration; however, these patterns differed slightly for each measurement of eye gaze, and for each group.

Percent fixations. In the first 500ms of sentence presentation, both groups had a greater percentage of fixations on the eyes than the mouth (audiovisual: M_{eyes} = 65%, SD = 21.5%; M_{mouth} = 33%, SD = 21.1%; $t(29) = 4.05$, $p < 0.001$, $d = 0.74$; audio-only: M_{eyes} = 71%, SD = 19.7%; M_{mouth} = 24%, SD = 18.4%; $t(28) = 6.57$, $p < 0.001$, d = 1.22). From 500ms onwards, the percentage of fixations increased on the mouth and decreased on the eyes in both groups, but this change was greater and lasted longer for the audiovisual group; indeed, we observed a significant interaction between IA and time, $F(2.04, 112.4) = 64.72$, $p < 0.001$, $\eta_p^2 = 0.54$, and an interaction between IA, time and group, $F(6, 330) = 37.19$, $p < 0.001$, $\eta_p^2 = 0.40$. For the audiovisual group, the mean percentage of fixations on the mouth significantly increased by a maximum of 50% from 500 ms – 2500 ms, and significantly decreased on the eyes by a maximum of 48% from 500 ms – 3000 ms (ps < 0.003). In the audio-only group, the mean percentage of fixations on the mouth significantly increased by a maximum of 10% from 500 ms – 1500 ms, and significantly decreased on the eyes by a maximum of 13% from 500 ms – 2500 ms (ps < 0.003). Overall, the audiovisual group had a greater percentage of fixations on the mouth than the eyes (M_{mouth} = 71%, SD = 23.2%; M_{eyes} = 28%, SD = 23.1%; $p < 0.001$) whereas the audio-only group had a greater percentage of fixations on the eyes than on the mouth (M_{eyes} = 62%, SD = 19.3%; M_{mouth} = 32%, SD = 18.8%; p < 0.001), revealed by a significant interaction between group and IA, $F(1,55) = 46.21$, p < 0.001, $\eta_p^2 = 0.46$.



Figures 3.1 – 3.4. Percentage fixations (Figs. 3.1 and 3.2) and mean fixation duration (Figs. 3.3 and 3.4) on the eyes and mouth during presentation of individual noise-vocoded sentences, for the audiovisual and audio-only groups. Time represents time from sentence onset. Error bars represent ± 1SE.

Fixation duration. In the audiovisual group, fixations on the mouth were overall significantly longer than fixations on the eyes (M_{mouth} = 1463 ms, SD = 615.9 ms; M_{eyes} = 516 ms, SD = 465.7 ms; $p < 0.008$), and at every time point ($ps < 0.007$). In the audio-only group, there was no significant difference in the duration of fixations on the eyes and mouth overall (M_{mouth} = 746 ms, SD = 395.8 ms; M_{eyes} = 847 ms, SD = 546.9 ms; $p > 0.008$), or at any time point ($ps > 0.007$). Fixations on the mouth in the audiovisual group were also significantly longer than fixations on the eyes, and on the mouth, in the audio-only group ($p < 0.008$). These patterns were confirmed by a significant IA x group interaction, $F(1,55) = 50.18$, $p < 0.001$, $\eta_p^2 = 0.48$.

Fixations in both groups significantly increased in duration during sentence presentation, although the amount and duration of this increase differed between groups and IAs. In the audiovisual group, fixations on the mouth significantly increased in duration by a maximum of 898 ms (SD = 488.9 ms) between 500 and 2500 ms, while fixations on the eyes increased in duration by a maximum of 283 ms (SD = 140.5 ms) between 500 and 3000 ms, $ps < 0.001$. In the audio-only group, fixations on the mouth significantly increased in duration by a maximum of 361 ms (SD = 346.5 ms) between

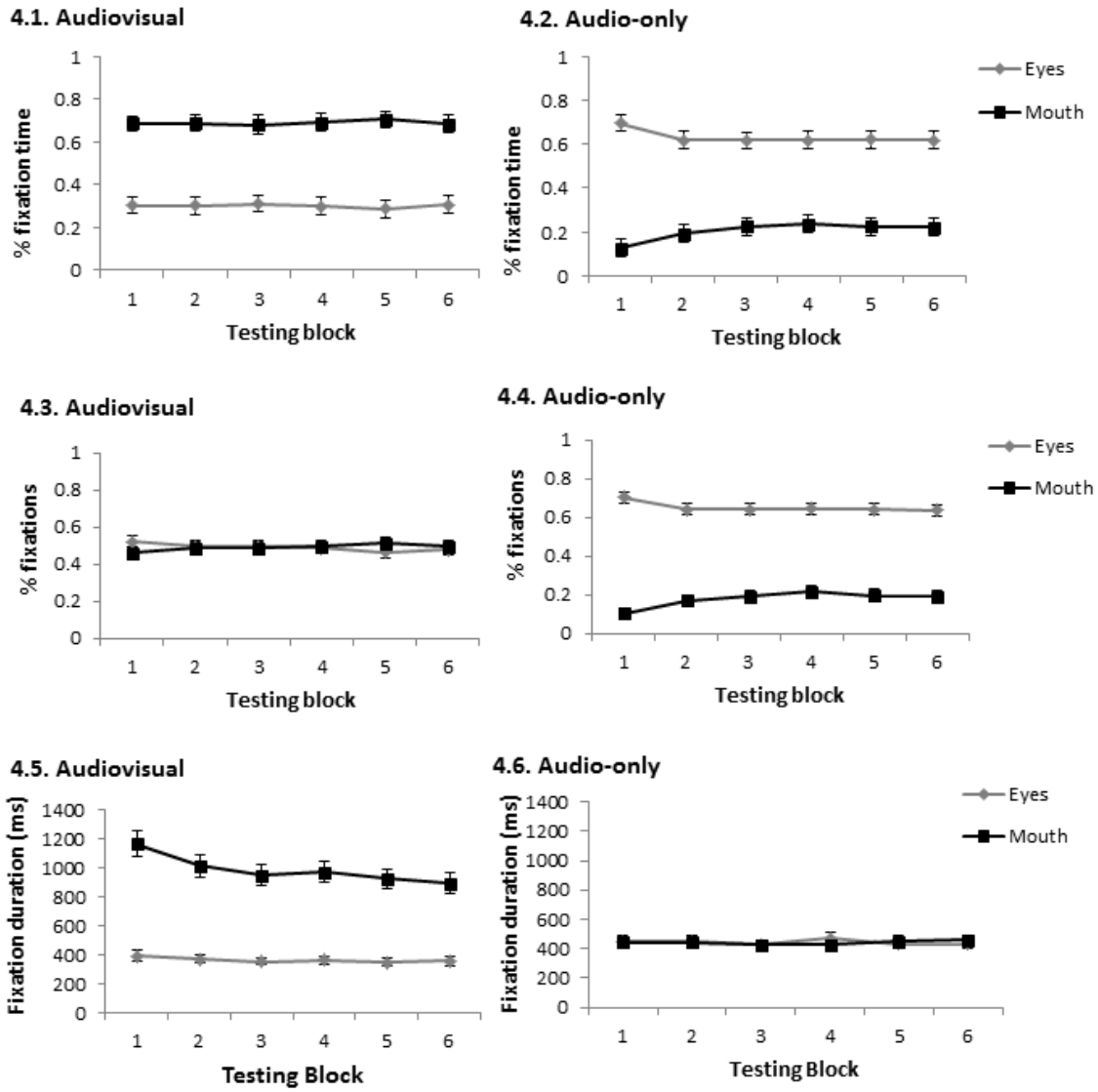
500 and 2500 ms, while fixations on the eyes increased in duration by a maximum of 459 ms (SD = 339.6 ms) between 500 and 3000 ms, *p*s < 0.001. These patterns were confirmed by a significant effect of time, $F(1.9,106.7) = 73.35$, $p < 0.001$, $\eta_p^2 = 0.57$, an interaction between time and IA, $F(2.2,118.37) = 8.16$, $p < 0.001$, $\eta_p^2 = 0.13$, and an interaction between time, IA and group, $F(6,330) = 12.54$, $p < 0.001$, $\eta_p^2 = 0.19$.

Eye gaze during perceptual adaptation to noise-vocoded speech.

General patterns. Figures 4.1 – 4.6 show data for each eye gaze variable on the speaker's mouth and eyes across the 6 testing blocks, for each group. In the audiovisual group, over 99% of all fixations fell on the speaker's face and 98% were on the eyes and mouth. In comparison, 83% of fixations from the audio-only group were on the speaker's face, and 74% on the eyes and mouth. Overall, perceivers in the audiovisual group looked more at the mouth than the eyes, with greater percent fixation time and longer fixations on the mouth. Conversely, perceivers in the audio-only group had greater percent fixation time and percent fixations on the eyes than on the mouth. Subtle changes in eye movements over time were also evident: in the audiovisual group, the duration of fixations on the mouth decreased, while in the audio-only group, percent fixations increased slightly on the mouth.

Where do listeners look? Overall, the audiovisual group looked more at the speaker's mouth than the eyes, whereas the audio-only group looked more at the eyes than the mouth; that is, we observed a significant interaction between IA and group for each measurement of eye gaze. In the audiovisual group, percent fixation time was significantly greater on the mouth than on the eyes overall (M_{mouth} = 69%, SD = 21.1%; M_{eyes} = 30%, SD = 20.8%), $t(29) = 5.05$, $p < 0.001$, $d = 0.92$, and in every testing block (*p*s < 0.008), while in the audio-only group, percent fixation time was significantly greater on the eyes than on the mouth (M_{eyes} = 63%, SD = 17.3%; M_{mouth} = 18%, SD = 12.8%), $t(28) = 7.04$, $p < 0.001$, $d = 1.31$; group x IA, $F(1, 57) = 69.04$, $p < 0.001$, $\eta_p^2 = 0.55$. Fixations on the mouth in the audiovisual group (M_{mouth} = 984 ms, *SD* = 370.2 ms) were significantly longer than fixations on the eyes overall (M_{eyes} = 363 ms, *SD* = 154.4 ms), $t(29) = 9.09$, $p < 0.001$, $d = 1.66$, and in every testing block (*p*s < 0.008), while in the audio-only group, the duration of fixations on the eyes and mouth were not significantly different, $p > 0.05$. Fixations on the mouth in the audiovisual group were

also significantly longer than fixations on the eyes, or on the mouth, in the audio-only group ($p < 0.008$); group x IA, $F(1, 57) = 70.84$, $p < 0.001$, $\eta_p^2 = 0.55$. Percent fixations on the eyes and mouth did not significantly differ in the audiovisual group, $p > 0.05$, while in the audio-only group, a higher percentage of fixations fell on the eyes than on the mouth overall ($M_{eyes} = 65\%$, $SD = 19.1\%$; $M_{mouth} = 18\%$, $SD = 14.2\%$), $t(28) = 8.09$, $p < 0.001$, $d = 1.50$, and in every testing block ($ps < 0.008$); group x IA, $F(1, 57) = 30.49$, $p < 0.001$, $\eta_p^2 = 0.35$.



Figures 4.1 – 4.6. Mean percent fixation time, percent fixations, and fixation duration on the mouth and eyes, per testing block and per group. Error bars show ±1SE.

Do patterns of eye gaze change over time? In the audiovisual group, fixations on the mouth significantly decreased in duration over time, by a maximum of 269 ms

(SD = 323.3 ms) between block 1 and block 6, $t(29) = 4.56$, $p < 0.001$, $d = 0.83$, while the duration of fixations on the eyes did not significantly change between any two testing blocks (ps > 0.003). In the audio-only group, the duration of fixations on the eyes and the mouth did not significantly change between testing blocks (ps > 0.003); indeed, the interaction between IA, testing block and group was significant, $F(5, 285) = 5.32$, $p < 0.001$, $\eta_p^2 = 0.09$.

No other measurements of eye gaze changed significantly over time; however, non-parametric analyses identified a significant effect of testing block in the audio-only group for percent fixations on the mouth, $\chi^2(5) = 16.88$, $p = 0.005$, $\phi = 0.76$. There was a significant increase of 7% based on median values (or 8% based on the mean; IQR = 1.4 – 6.5%), between block 1 and block 3, $z = 3.25$, $p = 0.001$, $r = 0.60$.

Part C. Is eye gaze related to recognition of audiovisual noise-vocoded speech?

Correlational analyses

Tables 1.1 and 1.2 show descriptive statistics and correlations between recognition accuracy and the three eye gaze variables on the mouth, per group, and at two time points (earlier and later trials). We observed two significant correlations in the audiovisual group which indicated that eye gaze towards the speaker's mouth was related to recognition accuracy: early recognition accuracy was positively correlated with early fixation duration, $r = 0.37$, $p = 0.046$, 95% CI [0.06, 0.63], and with later fixation duration, $r = 0.38$, $p = 0.037$, 95% CI [0.08, 0.64]; that is, longer fixations throughout the experiment were related to better recognition of the noise-vocoded speech in earlier trials. However, these correlations did not survive correction for multiple comparisons ($p > 0.008$). Equivalent correlations in the audio-only group were negative (longer fixations on the mouth were associated with *poorer* recognition), and were not significant: early recognition accuracy x early fixation duration, $r = -0.16$, $p = 0.409$, 95% CI [-0.46, 0.07]; early recognition accuracy x later fixation duration, $r = -0.27$, $p = 0.512$, 95% CI [-0.59, 0.07]. Furthermore, the correlations differed significantly between the two groups: early recognition accuracy x early fixation duration, $z = 2.00$, $p = 0.045$, $r = 0.26$; early recognition accuracy x later fixation duration, $z = 2.46$, $p = 0.014$, $r = 0.32$.

No other correlations between the eye gaze variables and recognition accuracy were significant, in either group. However, non-parametric analyses identified a significant positive correlation between later recognition accuracy and later percent fixation time, $r = 0.36$, $p = 0.049$, 95% CI [0.01, 0.63], indicating that better recognition accuracy was related to more time spent fixating on the speaker's mouth.

Between the 'predictor' variables, early and later recognition accuracy were positively correlated in the audiovisual group, $r = 0.80$, $p < 0.001$, 95% CI [0.61, 0.91], and in the audio-only group, $r = 0.70$, $p < 0.001$, 95% CI [0.44, 0.86]. These correlations were not significantly different, $z = 0.84$, $p = 0.400$, $r = 0.01$, indicating that better recognition in early blocks was related to better recognition in later blocks, regardless of the presentation modality. The eye gaze variables percent fixation time and percent fixations on the mouth were positively correlated with one another in both groups and for early and later trials (see Tables 1.1 and 1.2). Early trials for each eye gaze variable were also correlated with later trials. The correlations between eye gaze variables did not differ significantly between groups except for early and later fixation duration; these variables were more highly correlated in the audiovisual compared to the audio-only group (audiovisual: $r = 0.95$, $p < 0.001$, 95% CI [0.92, 0.98]; audio-only: $r = 0.72$, $p < .001$, 95% CI [0.35, 0.91], $z = 3.36$, $p = 0.001$, $r = 0.44$).

Fixation duration over time: good and poorer performers

Correlations between the duration of fixations on the mouth and recognition accuracy suggested that longer fixations were related to better performance in the audiovisual group. However, we also observed that fixations on the mouth in this group significantly decreased over time as performance improved; if the decrease in fixation duration was also related to improved performance (as we had hypothesised), we would have expected a negative correlation. To investigate these findings in more detail, we analysed recognition accuracy and the duration of fixations on the mouth in the audiovisual group over time, by participants' performance in the speech recognition task (good or poor).

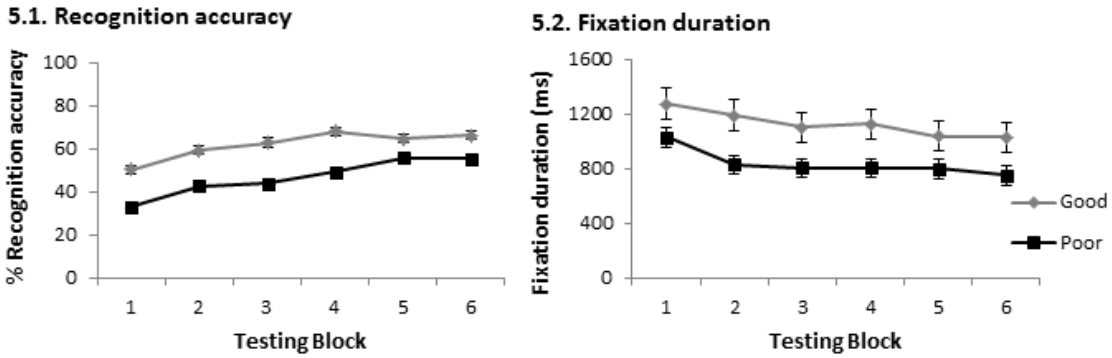
Table 1.1. Audiovisual group

Variable	M	SD	RA1	RA2	FT1	FT2	FD1	FD2	F1	F2
RA1 (%)	49	11.0	-	-	-	-	-	-		
RA2 (%)	60	12.4	0.80**	-	-	-	-	-		
FT1 (%)	69	21.1	0.33	0.32	-	-	-	-		
FT2 (%)	69	21.9	0.25	0.31	0.92**	-	-	-		
FD1 (ms)	1041	392.0	0.37*	0.27	0.43*	0.50*	-	-		
FD2 (ms)	928	357.2	0.38*	0.33	0.47*	0.56*	0.95**	-		
F1 (%)	48	16.9	0.26	0.27	0.92**	0.84**	0.29	0.26	-	
F2 (%)	50	17.7	0.20	0.27	0.86**	0.92**	0.30	0.30	0.92**	-

Table 1.2. Audio-only group

Variable	M	SD	RA1	RA2	FT1	FT2	FD1	FD2	F1	F2
RA1 (%)	31	9.1	-	-	-	-	-	-	-	-
RA2 (%)	39	9.4	0.70**	-	-	-	-	-	-	-
FT1 (%)	18	12.8	-0.04	-0.14	-	-	-	-	-	-
FT2 (%)	23	21.4	0.01	-0.07	0.81**	-	-	-	-	-
FD1 (ms)	441	173.5	-0.16	-0.30	0.46*	0.44*	-	-	-	-
FD2 (ms)	446	164.5	-0.27	-0.22	0.35	0.32	0.72**	-	-	-
F1 (%)	16	10.7	-0.08	-0.16	0.94**	0.74**	0.37	0.35	-	-
F2 (%)	20	19.4	0.00	-0.11	0.75**	0.94**	0.42*	0.30	0.77**	-

Tables 1.1 and 1.2. Correlation matrices for recognition accuracy and eye gaze on the mouth in the audiovisual (1.1) and audio-only (1.2) groups. All correlations are Pearson's *r*. RA = recognition accuracy; FT = percent fixation time; FD = fixation duration; F = percent fixations; '1' and '2' indicate early and later testing blocks. *$p < 0.05$; **$p < 0.001$.



Figures 5.1 and 5.2. Mean recognition accuracy (5.1) and mean duration of fixations on the mouth (5.2) for good and poorer performers in the audiovisual group, per testing block. Error bars show ±1SE.

Figures 5.1 and 5.2 show mean recognition accuracy and the duration of fixations on the mouth for good and poor performers in the audiovisual group. Analysis of recognition accuracy confirmed that good performers were significantly better than poor performers (M_{good} = 62%, SD = 9.1%; M_{poor} = 47%, SD = 7.2%), $F(1, 28) = 26.83$, $p < 0.001$, $\eta_p^2 = 0.49$, and that accuracy significantly increased in both groups, by a maximum of 18% (*SD* = 13.2%) between blocks 1 and 4 for good performers, and by a maximum of 22% (*SD* = 11.1%) between blocks 1 and 5 for poor performers, $F(4, 140) = 22.52$, $p < 0.001$, $\eta_p^2 = 0.45$. Analysis of fixation duration revealed that good performers had significantly longer fixations than poorer performers (M_{good} = 1130 ms, SD = 424.2 ms; M_{poor} = 839 ms, SD = 237.8 ms), $F(1, 28) = 5.36$, $p = 0.028$, $\eta_p^2 = 0.16$, as predicted by the positive correlation between these two variables. Fixations significantly decreased in duration for both good and poor performers, by a maximum of 244 ms (*SD* = 181.2 ms) for good performers, and by a maximum of 278 ms (*SD* = 440.0 ms) for poor performers, both between blocks 1 and 6, $F(2.3, 64.2) = 8.54$, $p < 0.001$, $\eta_p^2 = 0.23$. There was no interaction between time and performance ($p > 0.05$).

Discussion

The present study investigated patterns of eye gaze during recognition of, and perceptual adaptation to, audiovisual and audio-only noise-vocoded speech. The aim was to identify when listeners gain and make use of visual cues from the speaker's mouth, and whether eye gaze towards the mouth is related to performance. During recognition of individual audiovisual sentences, participants looked increasingly towards the mouth and fixations became significantly longer. As participants adapted to the audiovisual noise-vocoded speech, fixations on the mouth became significantly shorter. Furthermore, longer fixations on the mouth were related to better recognition of the noise-vocoded speech. Conversely, the audio-only group showed a general preference for looking at the speaker's eyes, and eye gaze was more stable over time; however, during recognition of individual sentences, fixations also became significantly longer, although less so than in the audiovisual group.

Part A. Is perceptual adaptation to noise-vocoded speech greater with audiovisual compared to audio-only cues?

As predicted, we observed a benefit of around 20% better recognition for the audiovisual group compared to the audio-only group, confirming that audiovisual cues improve recognition of noise-vocoded speech (Bernstein et al., 2013; Kawase et al., 2009; Pilling & Thomas, 2011). However, we did not replicate findings from these same studies that showed an audiovisual benefit for perceptual adaptation to noise-vocoded speech – that is, both groups adapted by equal amounts in the present experiment. Bayes factor indicated that our data were inconclusive rather than reflecting a null effect in the population, and this may have occurred for several reasons. Firstly, our participants gained a smaller benefit from the audiovisual cues in comparison to previous studies (for example, Pilling & Thomas (2011) observed a benefit of around 40%), and this may have been insufficient to improve perceptual adaptation. We also used different stimuli to previous studies; the IEEE sentences are longer and more complex than the Bamford-Kowal-Bench (BKB) sentences (Pilling & Thomas, 2011), single words (Kawase et al., 2009), or syllables (Bernstein et al., 2013), and may therefore have proved more difficult for participants to speech-read. Lastly, baseline recognition was not equated between groups, and was relatively high for the audiovisual group compared to previous studies (50% compared to 25% in Pilling & Thomas, 2011). This allowed us to compare speech recognition between the two groups, but consequently there was less room for improvement in the audiovisual group (that is, compared to previous studies and to the audio-only group). As listeners with lower starting accuracy tend to adapt to unfamiliar speech the most (Banks, Gowen, Munro, & Adank, 2015), the different baseline accuracy between the two groups may have masked any beneficial effects from the audiovisual cues.

Part B. When do perceivers of audiovisual noise-vocoded speech use visual speech cues?

Patterns of eye gaze changed significantly during individual sentence presentation. Within the first 1000 ms, eye gaze shifted from the eyes towards the mouth, replicating observations that listeners rapidly start to look towards the mouth following speech onset in quiet listening conditions (Lansing & McConkie, 2003).

Participants likely shifted their gaze to the mouth in order to compensate for the degraded auditory signal by making use of the visual speech cues – that is, changes in eye gaze were related to the task demands and the salience of the visual information. We predicted that visual cues would be most useful at the start of the sentence, when listeners start to 'tune in' to the unfamiliar speech and make predictions about the unfolding sentence. However, eye gaze towards the mouth peaked at around 2500 ms (over halfway through the sentences), when we observed the longest and greatest percentage of fixations on the mouth, suggesting that there is a delay in fully exploiting visual information from a speaker's mouth. This could be due to a delay in shifting eye gaze towards the visual speech cues – for example, a covert shift in attention can precede overt eye movements towards a target (e.g. Deubel & Schneider, 1996; Rizzolatti, Riggio, & Sheliga, 1994; Treisman & Gelade, 1980). It could also indicate that listeners first obtain the 'gist' of the visual scene (i.e. the speaker's face), before obtaining the more salient visual speech information slightly later (e.g. Loftus, 1981, 1983). The observed pattern therefore suggests that the timing of eye movements towards the mouth is not primarily driven by linguistic factors, but rather by attentional or oculomotor processes. This matches previous data showing that there was no relationship between fixations on a speaker's mouth and the phonetic content of the speech (Vatikiotis-Bateson et al., 1998).

During perceptual adaptation to the noise-vocoded speech (that is, over the course of the whole experiment), we observed a significant decrease in the duration of fixations over time – only on the mouth, and only in the audiovisual group. We had predicted that participants in the audiovisual group would look less at the mouth as they adapted to the noise-vocoded speech (that is, as their performance improved), and this observation matched our predictions. A decrease in fixation duration may reflect changes in eye gaze related to perceptual adaptation; that is, as perceivers' performance improved, they relied less on the visual speech cues, or were able to process them more efficiently. However, the observed decrease could also reflect a decrease in cognitive effort.

We predicted that eye gaze would remain stable over time in the audio-only group, as there was no task-related reason for them to modify their patterns of eye gaze; however, during recognition of individual sentences, fixations significantly increased in

duration until over halfway through the sentence (at 2500 – 3000 ms) as they had done in the audiovisual group, albeit by a smaller amount. This unexpected finding suggests that longer fixations did not solely relate to the processing of visual speech cues; in fact, they may have reflected overall cognitive effort involved in processing the auditory speech signal as well as the visual information. Indeed, measurements of eye gaze have been associated with increased mental and cognitive effort, including saccade amplitude (May, Kennedy, Williams, Dunlap, & Brannan, 1990) and longer fixations, for example in perception of low-frequency words in reading studies (see Rayner, 1998, for a review), or in flight simulations for pilots (De Rivecourt, Kuperus, Post, & Mulder, 2008).

Overall, the audiovisual group looked more at the speaker's mouth than the eyes. We observed longer fixations and a greater percentage of time fixating the mouth compared to the eyes over the course of the experiment, and longer and a greater percentage of fixations on the mouth than the eyes during individual sentence recognition. These patterns replicate previous observations that listeners fixate more on a speaker's mouth than the eyes in adverse listening conditions (Buchan et al., 2007; 2008; Lansing & McConkie, 2003; Vatikiotis-Bateson et al., 1998). Conversely, participants in the audio-only group overall spent a greater percentage of time fixating the speaker's eyes than the mouth, and displayed more fixations in this area over the course of the experiment, and during individual sentence recognition. This replicates observations that perceivers look primarily towards the eyes during perception of static faces (e.g. Birmingham & Kingstone, 2009; Lansing & McConkie, 2003), most likely for social reasons. Overall, fixations in the audio-only group were also significantly shorter than fixations on the mouth in the audiovisual group. For audio-only perceivers, the image contained no visual information that would be useful to the task, and they were thus likely 'scanning' the image of the speaker's face with no particular goal and shorter fixations. In contrast, the audiovisual group were likely attempting to decode the linguistically salient visual information as it unfolded temporally, with longer fixations on the mouth as a consequence.

Part C. Is eye gaze related to recognition of audiovisual noise-vocoded speech?

Better recognition of the noise-vocoded speech in earlier trials was related to longer fixations on the mouth in earlier and later trials, and only in the audiovisual group. This confirmed our hypothesis that eye gaze during recognition of unfamiliar audiovisual speech is related to performance, and suggests that participants who looked more steadily and for longer at the speaker's mouth performed better in the speech recognition task. As visual perception is reduced during eye movements (Matin, 1974), longer fixations on the mouth (and consequently fewer eye movements) likely reflect more effective decoding of visual speech cues. However, the correlations can be interpreted in two ways: 1) participants who are better at decoding visual speech cues look more at the speaker's mouth, or 2) participants who look more at the mouth are better able to decode the visual speech cues. Furthermore, the correlations did not survive correction for multiple comparisons. Further testing, either in a larger sample or by manipulating participants' eye gaze, could help to confirm the correlations reported here and constrain these possible interpretations.

To our knowledge, ours is the first study to demonstrate a relationship between eye gaze and recognition of audiovisual speech in adverse listening conditions. Previous studies have investigated the role of eye gaze towards a speaker's mouth, but have found no correlation with recognition of speech in noise (Buchan et al., 2007), or with speech in quiet (Everdell et al., 2009). In both of these studies, recognition accuracy was almost at ceiling (86% and 90% respectively), whereas in the present study it was much lower, possibly accounting for the different results reported here.

Lansing & McConkie (2003) investigated the relationship between eye gaze and speech-reading performance, observing that more gaze time on a speaker's mouth was very weakly related to poorer speech-reading (the correlation accounted for <1% of the variance). Their finding suggested that we were likely to observe a negative relationship between eye gaze on the mouth and recognition of noise-vocoded speech, and indeed this is what we had predicted. Furthermore, the positive correlations that we observed did not match our hypothesis that fixations decreased in duration over time because performance improved, as this would also predict a negative correlation (that is, shorter fixations in later trials would relate to better performance). To investigate this further,

we compared recognition accuracy and fixation duration over time for good and poorer performers in the speech recognition task, for the audiovisual group only. These analyses revealed that good performers had longer fixations on the mouth than the poorer performers, but that fixations on the mouth in both groups became shorter over time. This suggests that the relationship between eye gaze and performance is complex, and that the duration of fixations may reflect overall cognitive effort as well as processing of visual speech cues; that is, as participants' performance improved, or as they relaxed into the task, they expended less effort in decoding the noise-vocoded speech, and this was reflected in the decreasing duration of their fixations.

Conclusion

To our knowledge, the present study is the first to describe changes in eye gaze during recognition of individual noise-vocoded sentences, and during perceptual adaptation to noise-vocoded speech. The percentage and duration of fixations towards the speaker's mouth increased rapidly during audiovisual sentence recognition, but peaked over halfway through sentence presentation. This indicates that visual speech cues are used later than we predicted, and that the timing of eye gaze towards the speaker's mouth is driven by attentional and oculomotor processes rather than linguistic factors. The duration of fixations towards the speaker's mouth also decreased during perceptual adaptation to the noise-vocoded speech, but only in the audiovisual group, suggesting less reliance on visual speech cues as performance improved – that is, the decrease fixation duration was potentially driven by the listener's linguistic needs. We further demonstrated that longer fixations on the mouth in the audiovisual group were related to better recognition of the noise-vocoded speech, indicating that longer fixations are related to efficient processing of visual speech cues. However, fixations also significantly increased in duration during individual sentence recognition in the audio-only group. As there was no task-related reason for an increase in fixation duration, the pattern suggests that longer fixations may also be related to overall cognitive effort. Further investigation is needed to determine the exact meaning of longer fixations during recognition of unfamiliar audiovisual speech, and particularly the extent to which this reflects overall processing effort in comparison to processing of visual speech cues.

References

Banks, B., Gowen, E., Munro, K., & Adank, P. (2015). Cognitive predictors of perceptual adaptation to accented speech. *Journal of the Acoustical Society of America, 137*(4), 2015-2024. doi:10.1121/1.4916265

Bernstein, L. E., Auer, E. T., Jr., Eberhardt, S. P., & Jiang, J. (2013). Auditory perceptual learning for speech perception can be enhanced by audiovisual training. *Frontiers in Neuroscience, 7*. doi:10.3389/fnins.2013.00034

Birmingham, E., & Kingstone, A. (2009). Human Social Attention A New Look at Past, Present, and Future Investigations. *Year in Cognitive Neuroscience 2009, 1156*, 118-140. doi:10.1111/j.1749-6632.2009.04468.x

Boersma, P., & Weenink, D. (2012). Praat: doing phonetics by computer (Version 5.3.05).

Buchan, J. N., Pare, M., & Munhall, K. G. (2007). Spatial statistics of gaze fixations during dynamic face processing. *Social Neuroscience, 2*(1), 1-13. doi:10.1080/17470910601043644

Buchan, J. N., Pare, M., & Munhall, K. G. (2008). The effect of varying talker identity and listening conditions on gaze behavior during audiovisual speech perception. *Brain Research, 1242*, 162-171. doi:10.1016/j.brainres.2008.06.083

Christianson, S. A., Loftus, E. F., Hoffman, H., & Loftus, G. R. (1991). Eye fixations and memory for emotional events. *Journal of Experimental Psychology-Learning Memory and Cognition, 17*(4), 693-701. doi:10.1037//0278-7393.17.4.693

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., & McGettigan, C. (2005). Lexical information drives; Perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences. *Journal of Experimental Psychology-General, 134*(2), 222-241. doi:10.1037/0096-3445.134.2.222

De Rivecourt, M., Kuperus, M. N., Post, W. J., & Mulder, L. J. M. (2008). Cardiovascular and eye activity measures as indices for momentary changes in mental effort during simulated flight. *Ergonomics, 51*(9), 1295-1319. doi:10.1080/00140130802120267

Deubel, H., & Schneider, W. X. (1996). Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision Research, 36*(12), 1827-1837. doi:10.1016/0042-6989(95)00294-4

Dupoux, E., & Green, K. (1997). Perceptual adjustment to highly compressed speech: Effects of talker and rate changes. *Journal of Experimental Psychology-Human Perception and Performance, 23*(3), 914-927. doi:10.1037//0096-1523.23.3.914

Erber, N. P. (1975). Auditory-visual perception of speech. *Journal of Speech and Hearing Disorders, 40*(4), 481-492.

Everdell, I. T., Marsh, H., Yurick, M. D., Munhall, K. G., & Pare, M. (2007). Gaze behaviour in audiovisual speech perception: Asymmetrical distribution of face-directed fixations. *Perception, 36*(10), 1535-1545. doi:10.1068/p5852

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying fixed effects analyses of variance and covariance. *Review of Educational Research, 42*(3), 237-288. doi:10.3102/00346543042003237

Golomb, J. D., Peelle, J. E., & Wingfield, A. (2007). Effects of stimulus variability and adult aging on adaptation to time-compressed speech. *Journal of the Acoustical Society of America, 121*(3), 1701-1708. doi:10.1121/1.2436635

Grant, K. W., Walden, B. E., & Seitz, P. F. (1998). Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration. *Journal of the Acoustical Society of America, 103*(5), 2677-2690. doi:10.1121/1.422788

Greenwood, D. D. (1990). A cochlear frequency-position function for several species - 29 years later. *Journal of the Acoustical Society of America, 87*(6), 2592-2605. doi:10.1121/1.399052

Harwell, M. R., Rubinstein, E. N., Hayes, W. S., & Olds, C. C. (1992). Summarizing monte-carlo results in methodological research - the 1-factor and 2-factor fixed effects anova cases. *Journal of Educational Statistics, 17*(4), 315-339. doi:10.3102/10769986017004315

Henderson, J. M., Malcolm, G. L., & Schandl, C. (2009). Searching in the dark: Cognitive relevance drives attention in real-world scenes. *Psychonomic Bulletin & Review, 16*(5), 850-856. doi:10.3758/pbr.16.5.850

Hervais-Adelman, A., Davis, M. H., Johnsrude, I. S., & Carlyon, R. P. (2008). Perceptual learning of noise vocoded words: Effects of feedback and lexicality. *Journal of Experimental Psychology-Human Perception and Performance, 34*(2), 460-474. doi:10.1037/0096-1523.34.2.460

IEEE. (1969). Ieee recommended practice for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics, AU17*(3), 225-&.

Kawase, T., Sakamoto, S., Hori, Y., Maki, A., Suzuki, Y., & Kobayashi, T. (2009). Bimodal audio-visual training enhances auditory adaptation process. *NeuroReport, 20*(14), 1231-1234.

Langton, S. R. H., Watt, R. J., & Bruce, V. (2000). Do the eyes have it? Cues to the direction of social attention. *Trends in Cognitive Sciences, 4*(2), 50-59. doi:10.1016/s1364-6613(99)01436-9

Liversedge, S. P., & Findlay, J. M. (2000). Saccadic eye movements and cognition. *Trends in Cognitive Sciences, 4*(1), 6-14. doi:10.1016/s1364-6613(99)01418-7

Lix, L. M., Keselman, J. C., & Keselman, H. J. (1996). Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance F test. *Review of Educational Research, 66*(4), 579-619. doi:10.3102/00346543066004579

Loftus, G. R. (1981). Tachistoscopic simulations of eye fixations on pictures. *Journal of Experimental Psychology-Human Learning and Memory, 7*(5), 369-376.

Loftus, G. R. (1983). Eye fixations on text and scenes. In K. Rayner (Ed.), *Eye movements in reading: perceptual and language processes* (pp. 359-376). New York: Academic Press.

Macleod, A., & Summerfield, Q. (1987). Quantifying the contribution of vision to speech perception in noise. *British Journal of Audiology, 21*(2), 131-142. doi:10.3109/03005368709077786

Matin, E. (1974). Saccadic suppression - review and an analysis. *Psychological Bulletin, 81*(12), 899-917. doi:10.1037/h0037368

May, J. G., Kennedy, R. S., Williams, M. C., Dunlap, W. P., & Brannan, J. R. (1990). Eye-movement indexes of mental workload. *Acta Psychologica, 75*(1), 75-89. doi:10.1016/0001-6918(90)90067-p

Pilling, M., & Thomas, S. (2011). Audiovisual Cues and Perceptual Learning of Spectrally Distorted Speech. *Language and Speech, 54*, 487-497. doi:10.1177/0023830911404958

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*(3), 372-422. doi:10.1037/0033-2909.124.3.372

Rizzolatti, G., Riggio, L., & Sheliga, B. M. (1994). Space and selective attention. *Attention and Performance Xv: Conscious and Nonconscious Information Processing, 15*, 231-265.

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech Recognition with Primarily Temporal Cues. *Science, 270*(5234), 303-304.

Sommers, M. S., Tye-Murray, N., & Spehar, B. (2005). Auditory-visual speech perception and auditory-visual enhancement in normal-hearing younger and older adults. *Ear and Hearing, 26*(3), 263-275. doi:10.1097/00003446-200506000-00003

Sumby, W., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America, 26*(2), 212-215.

Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd & R. Campbell (Eds.), *Hearing by Eye: The Psychology of Lip-Reading*. Hillsdale, NJ: Lawrence Erlbaum.

Treisman, A. M., & Gelade, G. (1980). Feature-integration theory of attention. *Cognitive Psychology, 12*(1), 97-136. doi:10.1016/0010-0285(80)90005-5

Vatikiotis-Bateson, E., Eigsti, I. M., Yano, S., & Munhall, K. G. (1998). Eye movement of perceivers during audiovisual speech perception. *Perception & Psychophysics, 60*(6), 926-940. doi:10.3758/bf03211929

Vo, M. L. H., Smith, T. J., Mital, P. K., & Henderson, J. M. (2012). Do the eyes really have it? Dynamic allocation of attention when viewing moving faces. *Journal of Vision, 12*(13). doi:10.1167/12.13.3

CHAPTER 6.

GENERAL DISCUSSION

In this chapter, the novelty and impact of the experimental work conducted in each study is evaluated. The theoretical and methodological implications for each study are discussed, as well as specific limitations. Possible directions for future research are then outlined, as well as some general limitations to the work as a whole.

6.1 Cognitive predictors of perceptual adaptation to accented speech

Using an individual differences approach, this study demonstrated that a measure of executive function (inhibition) predicted perceptual adaptation to accented speech, while vocabulary knowledge predicted recognition of accented speech. Working memory had in indirect effect on recognition accuracy, mediated by vocabulary knowledge.

6.1.1 Novelty and impact of the work

6.1.1.1 Inhibition predicts perceptual adaptation to accented speech

To the author's knowledge, this is the first study to demonstrate a relationship between perceptual adaptation to unfamiliar speech and inhibition. Previous studies have demonstrated a link between perceptual adaptation and different measures of attention (Huyck & Johnsrude, 2012; Janse & Adank, 2012), but no study has specifically examined inhibitory abilities. Inhibition is a component of executive function and is related to attention, however it has been demonstrated that it is a separable component (Miyake et al., 2000). The finding therefore represents a step forward in understanding the cognitive mechanisms behind perceptual adaptation to accented speech, and specifically that being able to inhibit dominant or automatic behavioural responses helps listeners to improve their performance over time. Specifically, a new hypothesis is proposed – that perceptual adaptation to accented speech requires listeners to inhibit incorrect lexical responses triggered by hearing the

accented speech, based on a neighbourhood activation model (Luce & Pisoni, 1998). The finding therefore creates opportunities for future studies to test this hypothesis.

6.1.1.2 Vocabulary knowledge predicts recognition of accented speech, and mediates the relationship between working memory and recognition of accented speech.

To the author's knowledge, this is the first study to demonstrate a mediated relationship between vocabulary knowledge, working memory and recognition of accented speech using a path analysis model. This finding suggests that vocabulary knowledge is a more important cognitive ability than working memory for successful recognition of accented speech, and that working memory supports access to lexical knowledge rather than recognition of the accented speech per se. This is a somewhat controversial finding, as working memory has been proposed as vital to recognising speech in adverse listening conditions (Ronnberg et al., 2008), and is a reliable predictor of speech recognition in background noise (Akeroyd, 2008). Furthermore, Janse and Adank (2012) found working memory to predict recognition of accented speech in older adults even when vocabulary knowledge was taken into account. The role of working memory may therefore depend on the exact listening context, and the population in question, as other studies have found no relationship between working memory and recognition of accented (Gordon-Salant et al., 2013), frequency-compressed (Ellis & Munro, 2013) or noise-vocoded speech (Erb et al., 2012). A strength of the present study is the large sample size (N = 100), providing reliable data from a young, healthy population that is sufficiently powered to detect small effects. Further evidence from studies with a comparable sample size may help to clarify the role of working memory in relation to other cognitive abilities, as well as the exact contexts in which it is most important.

The relationship between vocabulary knowledge and recognition of accented speech adds to evidence that lexical information, and the ability to access and use lexical and semantic information, is highly important for successful recognition of accented speech (Davis et al., 2005; Janse & Adank, 2012; Loebach et al., 2010; Lively et al., 1994; Norris et al., 2003). The contribution of linguistic abilities such as vocabulary skill is still little researched, considering the amount of evidence suggesting

that lexical information is key to perceptual adaptation. Our results suggest that future studies should include similar measures to further investigate the role of linguistic skills in successful recognition of, and adaptation to, accented speech, and particularly how these skills interact with other cognitive abilities such as working memory.

6.1.1.3 Perceptual adaptation to accented speech, and recognition of accented speech, involve different cognitive mechanisms

The research reported in Chapter 3 contributes to an overall understanding of the mechanisms behind recognition of, and adaptation to, accented speech, highlighting the key role of cognition but also the complex interrelationships between different abilities. Particularly, different cognitive predictors for perceptual adaptation to, and recognition of the accented speech, were identified. This suggests that these two measurements of perceptual plasticity (that is, how listeners respond to adverse listening conditions) are in fact separate processes, driven by different cognitive abilities. Our results highlight the importance of considering both of these measures when investigating perceptual plasticity in relation to unfamiliar speech; specifically, that an individual's overall ability, and their ability to learn, are two separate processes, and should be considered as such in the literature.

6.1.1.4 Using path analysis to build a comprehensive model of perceptual plasticity

The study presents a novel way of testing the contribution of different cognitive abilities in the context of speech recognition and perceptual adaptation, using path analysis. This method is used widely in the social sciences (see Wolfle, 2003 for a review), and has, for example, been used to assess different aspects of cognition in visual perceptual adaptation (Kennedy, Rodrigue, Head, Gunning-Dixon, & Raz, 2009). To the author's knowledge, this is the first study to use this method in the context of speech recognition in adverse listening conditions. Research in this area has tended to focus on individual aspects of cognition such as working memory, but the research reported here highlights the need for a comprehensive model, taking into account multiple abilities that contribute to overall recognition and adaptation, and the complex relationships between them. Although path analysis is not suitable for exploratory work, the method may prove particularly useful in building such a model based on existing

evidence – for example, a single model could be used to test the relative contribution and interactions between cognitive and sensory abilities, that are already known to contribute to speech recognition or perceptual adaptation in adverse conditions.

6.1.2 Limitations of the work

6.1.2.1 Neuropsychological tests as a measure of cognitive ability

As discussed in the General Methods, there are some inherent limitations to using neuropsychological tests as a measure of cognition. Primarily, neuropsychological tests are an indirect measure of cognition, and may therefore tap into more than one cognitive ability. For example, it is possible that the measure of inhibition used in the study also indicates an individual's overall focus or attention on the task, and that the relationship between inhibition and perceptual adaptation reflects this – that is, participants who were better focused on both tasks performed better. Similarly, vocabulary knowledge is highly correlated with IQ (Kamphaus, 2005; Wechsler, 1958), and the relationship that was observed between vocabulary knowledge and recognition of accented speech, could therefore reflect better overall IQ. Vocabulary knowledge was also correlated with working memory, and indeed, the tests used for these measurements rely somewhat on overlapping abilities (accurate mapping between lexical items and semantic concepts); this may explain the mediation effect present between them in this study. However, these limitations are not unique to the present research. Neuropsychological tests have been used widely in similar research (e.g. Erb et al., 2012; Janse, 2009; or see Akeroyd, 2008, for a review), and they provide a simple and effective way of measuring cognitive ability in an experimental setting, providing important evidence of the cognitive processes likely involved in perceptual plasticity. Nevertheless, it would be beneficial to replicate the results reported here in future research, for example using experimental manipulations to provide a tighter control over measurements of cognition.

6.1.2.2 Other potential predictors of perceptual plasticity

Three measures of cognitive ability that contribute to recognition of, and perceptual adaptation to, accented speech, were identified. Even so, they only accounted for a relatively small proportion of the overall variation. This implies that other abilities,

not measured in the present study, also significantly contribute to individual performance, and highlights the need for a comprehensive model taking into account all of the mechanisms behind recognition and adaptation. Other potential cognitive predictors of perceptual plasticity of unfamiliar speech are discussed in the following section (6.1.3).

6.1.3 Future Research

Much work still needs to be done in order to fully understand the cognitive mechanisms of perceptual plasticity in relation to unfamiliar speech. Firstly, the role of executive function needs to be investigated more thoroughly. There is growing evidence that executive function contributes to perceptual adaptation to unfamiliar speech (e.g. Huyck & Johnsrude; Janse & Adank, 2012); however, we still do not fully understand how, or which components of executive function, contribute to perceptual adaptation. We have identified that inhibition may play a role in adaptation to accented speech, but other components, such as information updating and monitoring, may also contribute. Furthermore, the study reported here presents a hypothesis that perceptual adaptation involves inhibiting incorrectly activated lexical responses triggered by the accented speech. This interpretation could be tested using a visual world paradigm and eye tracking, whereby participants are simultaneously presented with an (auditory) accented sentence, and several pictures on screen which correspond to lexical neighbours that might potentially be triggered by the accented speech, as well as the correct lexical item. If the hypothesis is correct, participants who look less towards the distractor items would adapt the most. However, the observed correlation between perceptual adaptation and inhibition may also be related to overall attention during the tasks. Testing different aspects of listeners' attention, including inhibition, in a confirmatory factor analysis, could confirm which components of attention contribute to perceptual adaptation, and to what extent inhibition contributes independently of other attentional processes.

Secondly, the role of vocabulary knowledge, lexical processing and other linguistic skills is an interesting and worthwhile area of investigation, given the mounting evidence that lexical processing is key to perceptual adaptation of unfamiliar speech (Davis et al., 2005; Janse & Adank, 2012; Loebach et al., 2010; Lively et al., 1994; Norris et al., 2003). A possible interpretation of the correlation between

vocabulary knowledge and recognition of accented speech, is that better vocabulary skill specifically helps to *predict* the unfolding unfamiliar sentence by allowing listeners to easily access lexical items, thus relying on accurate and strong mappings between lexical items and semantic concepts. This hypothesis could be tested in future research by correlating participants' vocabulary skill with recognition of easily predictable compared to unpredictable sentences. Alternatively, vocabulary skill could be correlated with anticipatory eye movements towards target objects on screen presented simultaneously with the (auditory) accented sentence; this method has previously been used to identify a link between the prediction of upcoming lexical items (for clear, familiar speech) with vocabulary knowledge (Borovsky et al., 2012), and working memory (Huettig & Janse, 2015).

Furthermore, the relationship between vocabulary skill and working memory needs to be clarified; particularly, was the observed mediation effect a product of the true relationship between working memory, vocabulary skill and recognition of accented speech, or was it caused by an overlap in the tests used? To clarify this, a further correlational or confirmatory factor analysis could be carried out employing a different working memory test that does not rely on lexical and semantic processing – e.g. a digit span or phonological working memory test.

Thirdly, the remaining predictors of recognition of, and perceptual adaptation to, accented speech, need to be identified. As stated earlier in this chapter, the present studies have only accounted for a small proportion of the variance in perceptual adaptation and recognition of accented speech, and more work is thus required to identify additional predictors to construct a comprehensive model of both abilities. This could involve pattern or statistical learning (Neger, 2013), measures of sensory ability, cognitive effort (Zekveld et al., 2010), or overall intelligence (Amitay et al., 2010). In the present research, participants with lower baseline recognition also adapted the most, implying that a degree of motivation (that is, the motivation to perform better in the task) drives individuals to adapt, as has been shown with auditory perceptual learning (Amitay et al., 2010). This could be explored through the use of gaming techniques or providing feedback, for example. Other social and emotional factors could also play a role, for example the relationship between the listener and speaker, or the dynamics of turn-taking during conversation, can affect speech recognition, particularly in older

adults (Pichora-Fuller, 2003). Increased levels of anxiety in the listener have also been shown to reduce the accuracy of phoneme discrimination (Mattys, Seymour, Attwood, & Munafo, 2013). Studying the interaction between social and emotional factors and cognition, in the context of perceptual adaptation in adverse listening conditions, could provide an interesting and informative line of enquiry for future studies.

A particularly useful addition to the literature would also comprise experimental manipulations of cognitive ability to verify causation, as the majority of studies use a correlational design. Furthermore, the work here only examines adaptation to accented speech, and we do not know if these results generalise to other types of unfamiliar speech, such as acoustic distortions. Testing different aspects of cognition along with other possible predictors, in different types of unfamiliar speech, will help to build a comprehensive model for recognition of, and perceptual adaptation, to unfamiliar speech.

6.2 Audiovisual cues benefit recognition of accented speech in noise but not perceptual adaptation

Results from two experiments (Study 1 and Study 2) revealed that the availability of audiovisual speech cues does not lead to greater perceptual adaptation to accented speech, when compared to audio-only cues. Recognition of the accented speech in noise was greater with audiovisual cues than audio-only cues.

6.2.1 Novelty and impact of the work

6.2.1.1 Audiovisual cues do not improve perceptual adaptation to accented speech in young, healthy adults

The studies reported in Chapter 4 demonstrate that perceptual adaptation to accented speech is not improved by the presence of audiovisual cues in young, healthy adults; particularly, Bayesian analyses supported the null hypothesis. Potential confounds from a similar study (Janse & Adank, 2012) were addressed, by using eye tracking (to confirm that participants in the audiovisual group looked primarily at the speaker's face), and by testing a young, normal-hearing population. The hypothesis was

tested with two different accents and speakers, to ensure that the null result was not due to speaker- or accent-specific characteristics, and using two different experimental designs (training vs. continuous exposure to audiovisual speech). The results therefore help to constrain theories regarding the mechanisms of perceptual adaptation to accented speech; namely, that strategies improving overall perception of unfamiliar speech do not necessarily help listeners to learn. The benefits of learning unfamiliar speech with audiovisual cues may also be restricted to particular phonetic contrasts (e.g. Hazan et al., 2010), rather than accented speech per se.

6.2.1.2 Audiovisual speech cues benefit recognition of accented speech in noise

Both studies replicated previous results indicating that audiovisual cues are beneficial to recognition of accented speech in noise (Arnold & Hill, 2001; Janse & Adank, 2012; Kawase et al., 2014; Yi et al., 2013). This adds to existing evidence of the benefits of audiovisual cues to speech recognition in adverse listening conditions; however, as this was not the primary aim of the research, questions remain regarding the exact benefits of audiovisual cues to recognition of *accented* speech, as opposed to compensating for the background noise that our stimuli were presented in (see section 6.4.3).

The results reported here do, however, demonstrate that the benefits obtained from audiovisual cues vary greatly, as a different amount of benefit was observed between the two experiments. This could be due to differences in the novel and Japanese accent that we used (Kawase et al., 2014), or differences in the speakers (Kricos & Lesner, 1982, 1985). Although it was not an aim of the present research to investigate how different listening contexts affect the amount of benefit gained from audiovisual speech cues, it is nevertheless an interesting observation, and further research may help to identify the listening contexts for which audiovisual speech is most beneficial; for example, why some speakers afford greater benefits from their visual speech than others.

6.2.2 Limitations

6.2.2.1 The training design used in Study 1

Study 1 employed a training design, whereby participants were exposed to a period of training with audiovisual, audio-only or visual speech cues, in between testing sessions carried out in the auditory modality. However, results showed that there was no significant difference in perceptual adaptation between the control group (who underwent training with visual-only cues), and all other groups. This indicated that the training itself did not work. This could potentially be due to the duration and timing of the training, or because training conditions were inconsistent with the testing sessions (that is, stimuli in the training session were presented in varying levels of background noise in an adaptive procedure).

Nevertheless, training designs similar to ours have been used successfully to demonstrate that audiovisual cues can benefit perceptual adaptation to noise-vocoded speech (Bernstein et al., 2013; Kawase et al., 2009; Pilling & Thomas, 2011). It is thus unclear exactly why the design did not work. However, this limitation was addressed by retesting the hypothesis with a modified design in Study 2, whereby participants were continuously exposed to accented speech in either modality, thus eliminating any concerns about the training design affecting perceptual adaptation.

6.2.2.2 Different measurements of speech recognition were used in Study 1 and 2

To measure recognition accuracy, Speech Reception Thresholds (SRTs), gained through an adaptive procedure, were used in Study 2.1, while percentage recognition was used in Study 2.2. This was mainly due to technical limitations – that is, the adaptive staircase procedure was not easily adaptable to an audiovisual format. Consequently, comparisons between recognition accuracy and the amount and rate of perceptual adaptation cannot easily be made between studies. However, this was not a particular aim of the research, and large differences are unlikely between two samples of young, healthy adults drawn from the same population. Furthermore, amount of perceptual adaptation does not necessarily differ between accents (Pinet et al., 2011), and so the different measurements used do not constitute a major limitation of the study.

6.2.3 Future Research

Several questions remain unanswered regarding the benefits of audiovisual cues in perceptual adaptation to unfamiliar speech. Particularly, the results reported in Chapter 4 reveal that the benefits of audiovisual speech cues vary, perhaps between individuals as well as different contexts, and importantly that they are not always beneficial. Future research may focus on establishing the exact parameters of the benefits that can be obtained from audiovisual speech, for example directly comparing audiovisual speech recognition in different types of adverse listening conditions, and in different speakers. It would also be useful to establish whether they have a practical application, for example in clinical or aging populations, or even in second language learning (e.g. Hazan et al., 2010). An interesting direction to pursue may be an individual differences approach to audiovisual speech recognition, for example identifying the cognitive predictors of successful audiovisual speech recognition, or perceptual adaptation, in comparison to audio-only. Benefitting from audiovisual speech cues relies on successful speech-reading, and indeed speech-reading ability predicts the benefits obtained from audiovisual speech cues (Sommers et al., 2005). Understanding the factors behind individual variation in speech-reading (for example, the contribution of linguistic skills and experience, or other cognitive abilities such as working memory) may therefore also explain individual variation in the benefits obtained from audiovisual speech cues.

Lastly, we were not able to establish whether audiovisual speech cues specifically benefit recognition of accented speech, above and beyond their compensatory benefit for speech in background noise. Further research is therefore required to address this question. This would involve developing a task or stimuli that avoided ceiling effects following a short period of adaptation. The greatest potential for this may be in using a more complex novel accent, as the relative difficulty of such stimuli is easy to control.

6.3 Eye gaze during recognition of audiovisual noise-vocoded speech

This study investigated when visual speech cues are used during recognition of audiovisual noise-vocoded speech by measuring perceivers' eye gaze towards the speaker's mouth over time. The proportion and duration of fixations on a speaker's mouth increased during perception of audiovisual noise-vocoded sentences, while the duration of fixations decreased during a longer period of perceptual adaptation to the noise-vocoded speech. Longer fixations on the speaker's mouth were related to better recognition of audiovisual noise-vocoded speech. Longer fixations on the mouth may therefore indicate more effective processing of visual speech cues, but could also be related to overall cognitive effort. Finally, audiovisual cues improved recognition of the noise-vocoded speech, but not perceptual adaptation.

6.3.1 Novelty and impact of the work

6.3.1.1 Eye gaze towards a speaker's eyes and mouth varies during recognition of audiovisual noise-vocoded sentences, and during perceptual adaptation to noise-vocoded speech.

Individual sentences. Audiovisual perceivers rapidly shifted their gaze from the speaker's eyes to the mouth, fixating more often and for longer on the mouth as the sentence unfolded; conversely, audio-only perceivers looked consistently more at the eyes, demonstrating that shifts in eye gaze are driven by the saliency of visual information in relation to the task. For the audiovisual group, the proportion and duration of fixations on the mouth peaked over halfway through sentence presentation, indicating that the timing of eye movements is likely driven by attentional or oculomotor processes rather than linguistic; that is, the observed pattern did not match the hypothesis that listeners would look more at the mouth at the start of the sentence, in order to adapt to the unfamiliar speech, or to predict the upcoming sentence. Interestingly, an increase in the duration of fixations in both groups further suggested that longer fixations did not just reflect processing of visual speech cues, but also overall cognitive effort.

Perceptual adaptation. During perceptual adaptation to noise-vocoded speech (that is, as listeners' performance improved during presentation of multiple sentences),

the duration of fixations towards the speakers' mouth decreased, but only in the audiovisual group. This decrease was hypothesised to relate to adaptation to the noise-vocoded speech; that is, as listeners' performance improved, they relied less on the visual cues to decode the unfamiliar speech. This would suggest that variation in eye gaze over time, and use of visual speech cues, is goal-driven; that is, perceivers vary their eye gaze according to their needs during the task. However, the decrease in fixation duration could also reflect a decrease in overall effort.

To the author's knowledge, this is the first study to describe patterns of eye gaze during recognition of audiovisual noise-vocoded sentences, or during perceptual adaptation to unfamiliar speech. Results suggest that use of visual speech cues is dynamic and goal-driven, even over short periods of time, but that the timing of eye gaze to linguistically salient areas is not necessarily driven by linguistic needs. The data add to a large body of research into audiovisual speech perception in adverse listening conditions, including studies showing that eye gaze towards a speaker's eyes and mouth varies depending on the listening conditions, and the specific demands of the task (Buchan et al., 2007; 2008; Lansing & McConkie, 2003; Mital & Henderson, 2012; Vatikiotis-Bateson et al., 1998). However, the results presented here are novel as they specifically identify *when* listeners make use of speech cues from a speaker's mouth, when the task and listening conditions are held constant. They demonstrate that eye gaze is not just driven by environmental factors such as background noise – in fact, how and when visual cues are processed is likely determined by processes internal to the listener, such as the success of their auditory or visual perception, their reliance on the visual cues, or even their overall cognitive load or effort. The results also demonstrated that a measurement of eye gaze – fixation duration – may provide insight into cognitive effort during audiovisual *and* auditory speech processing. Measurements of eye gaze have been related to increased mental load during an auditory memory task (May et al., 1990), and fixation duration in particular has been used in studies investigating mental effort during tasks with a high cognitive load such as piloting aeroplanes (De Rivecourt et al., 2008) or anaesthesia (Schulz et al., 2011). However, fixation duration (or other measures of eye gaze) has not been used for research into speech or auditory perception, and may therefore be a useful compliment to existing techniques such as pupillometry,

particularly in paradigms where eye movements cannot be supressed (which is a requirement for accurate pupillometry measures).

6.3.1.2 Eye gaze is related to successful recognition of noise-vocoded speech.

Longer fixations on the speaker's mouth predicted better recognition of the noise-vocoded speech. To the author's knowledge, this is the first study to demonstrate such a relationship. Previous studies have investigated correlations between measurements of eye gaze (including the duration of fixations) and recognition of speech in noise (Buchan et al., 2007; Everdell et al., 2009; Lansing & McConkie, 2003) but have failed to observe a substantial correlation, possibly due to speech recognition being at ceiling level (Lansing & McConkie, 2003, showed a significant correlation but it accounted for <1 % of the variance in recognition accuracy). The present finding therefore confirms that eye gaze plays a role in recognition of unfamiliar audiovisual speech recognition. Particularly, steadily fixating a speaker's mouth area is important for successful recognition of unfamiliar speech. This provides an interesting basis for further research into strategies for improving recognition of unfamiliar speech, and particularly for speech-reading. This was previously suggested in a single case study by Lansing & McConkie (1999), which examined eye gaze patterns of a proficient speech-reader. The authors concluded that examining eye gaze in this context may be used to understand effective speech-reading, but to our knowledge this was not followed up.

However, it is not yet know if this particular behavioural strategy (that is, longer fixations on the mouth) drives better recognition, or if better recognition (or better speech-reading ability) drives the behaviour. If the former is true, training listeners to steadily fixate a speaker's mouth could potentially help individuals with a hearing impairment, or users of hearing aids, to improve their speech recognition in adverse listening conditions, for example in speech-reading classes. Furthermore, it is not known whether longer fixations on the speaker's mouth reflect successful processing of visual speech cues, or increased cognitive effort (as discussed in preceding paragraphs). Nevertheless, observing a correlation between fixations on the mouth and speech recognition strengthens our results regarding the timing of eye gaze to the speaker's mouth during audiovisual speech perception, as it indicates that looking at the mouth is in some way involved in recognition of audiovisual speech. This is a strength of the

present study compared to previous research, which has observed more and longer fixations towards the mouth in adverse conditions, but no relationship with recognition (Buchan et al., 2007; Everdell et al., 2009; Lansing & McConkie, 2003).

6.3.1.3 Eye gaze is consistently directed more towards the mouth than the eyes during perception of audiovisual noise-vocoded speech

Previous studies have demonstrated that perceivers of audiovisual speech look more towards a speaker's mouth than their eyes when background noise is present (Buchan et al., 2007; 2008; Lansing & McConkie, 2003; Vatikiotis-Bateson et al., 1998). We have extended this finding to perception of noise-vocoded speech; the audiovisual perceivers in our study looked consistently more at the speaker's mouth than the eyes, in comparison to audio-only perceivers who looked primarily at the speaker's eyes. The finding adds to evidence that perceivers look towards a speaker's mouth in adverse listening conditions to gain useful linguistic information in order to successfully recognise the unfamiliar speech.

6.3.1.4 Audiovisual cues do not always improve perceptual adaptation to noise-vocoded speech

Previous studies have clearly shown that audiovisual speech cues can improve perceptual adaptation to noise-vocoded speech (Bernstein et al., 2011; Kawase et al., 2009; Pilling & Thomas, 2011). However, this finding was not replicated in the present study, and a Bayes factor analysis indicated that our data were inconclusive. This could be due to sampling error, differences in the particular speaker or stimuli that we used (in comparison to previous studies), or the different levels of baseline accuracy between the two groups (see Chapter 5, Discussion for details). Nevertheless, the present data suggest that there is some individual variation in benefits from audiovisual speech during perceptual adaptation to unfamiliar speech. Furthermore, the effect may be sensitive to changes in baseline accuracy or the type of speech stimuli – that is, it is not a robust effect. The present study did, however, replicate the effect of better overall recognition of noise-vocoded speech with audiovisual cues compared to audio-only (Kawase et al., 2009; Pilling & Thomas, 2011), thus demonstrating the benefits that can be obtained from the presence of visual speech cues. Nevertheless, further testing is

required to clarify the exact parameters of the effect in relation to perceptual adaptation to noise-vocoded speech.

6.3.2 Limitations

6.3.2.1 Is eye tracking a reliable method for investigating use of visual speech cues?

Eye tracking has been used in several studies of audiovisual speech perception to investigate whether listeners look towards particular facial features in adverse conditions. Whether listeners look towards the speaker's mouth is of particular interest, as this is the most salient area for gaining visual linguistic information (Summerfield, 1987). A problem with using eye gaze in this context is that we cannot be certain whether looking towards a speaker's mouth indicates that the perceiver is processing the visual speech cues. Data from the present study, and particularly increases in the duration of fixations observed in the audio-only group, revealed that this assumption is not always correct. Indeed, longer fixations have been used as an indicator of a perceiver's mental effort (De Rivecourt et al., 2008; Schulz et al., 2011), and measurements of eye gaze could therefore indicate other cognitive processes not necessarily related to visual processing.

This has implications for the use of eye tracking in the context of audiovisual speech perception. It indicates that any conclusions drawn from the method must be done so with caution, and that experiments should be well controlled to make the correct inferences possible, for example including a control condition where visual information is redundant to the task such as the one in the present study. It also creates an interesting direction for future research, to investigate in more detail the aspects of cognition that longer fixations reflect during audiovisual speech perception, for example whether this measurement can indicate auditory processing effort as well as visual processing.

There are also some inherent limitations to using eye gaze (as measured by an eye tracker) as a measurement of cognitive processing. Eye-tracking measures the foveal visual field, but we cannot always guarantee that this is where a perceiver's attention is directed compared to, for example, their peripheral vision (Findlay &

Liversedge, 2000). In audiovisual speech perception, attention weighted towards the auditory and visual modalities can also vary (Hazan et al., 2010), which would not necessarily be picked up by an eye-tracker (for example, a listener may fixate the speaker's mouth but primarily attend to the auditory signal). Nevertheless, eye tracking is a relatively unexplored method for investigating audiovisual speech perception in adverse listening conditions, and could potentially be used to further explore different eye gaze strategies, for example in highly experienced speech readers (e.g. Lansing & McConkie, 1999), or even for assessing cognitive effort in auditory processing.

6.3.2.2 Using static images of the speaker's face as a control condition

The present study included a control group in order to determine whether patterns of eye gaze during audiovisual speech perception were unique to that particular context. The control condition comprised perception of audio-only speech, and static images of the speaker's face taken from the audiovisual recordings. This controlled for the presence of a facial image and the speaker's visual appearance, whilst varying the presence of linguistic visual information that was salient to the task. However, previous research has demonstrated that fixations are fewer and longer when viewing a dynamic (speaking) face compared to a static face, even in quiet listening conditions (Lansing & McConkie, 2003), and this may therefore have introduced a confound to our study. However, as a correlation between longer fixations on the mouth and speech recognition was also observed in the audiovisual group, it is unlikely that longer fixations were solely due to the dynamic nature of the stimuli.

6.3.2.3 Interest areas

A limitation of the interest areas in the present study is that they were not dynamic to match the speaker's moving face, and this could potentially have affected the accuracy of our analyses. However, several steps were taken to ensure that the interest area analyses were accurate: 1) the speaker was asked to keep their head completely still during recordings; 2) interest areas were individually set for each video clip; and 3) the interest areas were created to cover the whole eye and mouth area during any movements – that is, they were broad, but not overlapping, semi-circular areas around the eyes and the mouth. Broad interest areas were suitable for the purposes of the study, as we did not require a fine spatial analysis of eye gaze, but were interested in

whether participants were looking towards the lower or upper parts of the speaker's face (that is, towards the eyes or towards the mouth). Static interest areas, following these same steps, have successfully been used in previous studies of audiovisual speech perception (Everdell et al., 2007; Lansing & McConkie, 2003; Vatikiotis-Bateson et al., 1998).

A second potential limitation of the interest areas is that the speaker's nose was not included. Anecdotal reports suggest that skilled speech-readers specifically look more at a speaker's nose in order to speech-read effectively, but to our knowledge, this has not been documented in the scientific literature. Furthermore, a single case study of a skilled speech-reader demonstrated that she looked primarily between the speaker's mouth and eyes (Lansing & McConkie, 1999), with similar patterns to non-skilled speech-readers (Lansing & McConkie, 2003). As there was no specific reason for participants in the present study to look at the speaker's nose (that is, there was no social or linguistic reason for directing their gaze to this area, and they were not skilled speech-readers), it was not included as a specific interest area. Whilst it would be interesting to conduct a more detailed spatial analysis of eye gaze during audiovisual speech perception (e.g. Lansing & McConkie, 2003), this was not a particular aim of this study.

6.3.3 Future Research

Several interesting questions have been generated from the research presented in Chapter 5. Firstly, does fixating a speaker's mouth lead to better recognition of unfamiliar speech, or does better recognition influence the location of perceivers' eye gaze? Either hypothesis is possible; looking towards the mouth clearly provides useful linguistic cues, but this behaviour could be modified by a perceiver's ability to decode them. This could be tested simply by manipulating instructions to participants in an audiovisual speech recognition task, requiring them to either fixate the eyes or the mouth of the speaker only. Alternatively, eye gaze could be analysed in skilled speech-readers compared to unskilled speech-readers. Indeed, anecdotal reports suggest that speech-readers use particular strategies of eye gaze, particularly looking at the middle of the speaker's face. However, this behaviour, and the effectiveness of particular eye gaze strategies, have not been tested empirically (although, for a single case study, see

Lansing & McConkie, 1999). Understanding how different patterns of eye gaze relate to speech-reading performance may help develop strategies for certain individuals or populations who have difficulty understanding speech in adverse conditions, to better exploit the potential of audiovisual speech cues.

Secondly, the results implied that the duration of fixations (regardless of where they were directed on a speaker's face) could indicate overall cognitive processing effort; particularly, when no salient speech information is available, longer fixations may indicate increased cognitive effort in relation to auditory processing. This unexpected result merits further investigation. An experiment comparing fixation duration and pupil size (an indicator of listening effort; for example, see Zekveld et al., 2010) in a listening task with a static image, similar to the control condition in Chapter 5, could shed some light on whether longer fixations do indeed reflect effortful auditory processing (it should be noted that relative pupil size could not be analysed from the present data due to the lack of a suitable visual baseline). The duration of fixations have been used as an objective physiological indicator of mental load in applied psychological research, for example to assess the mental workload and performance of pilots (De Rivecourt et al., 2008) or anaesthetists (Schulz et al., 2011), but to the author's knowledge, this measure has not been used in speech perception or audiological research. Establishing exactly what longer fixations reflect during auditory processing could potentially lead to a measure of cognitive effort or load that will complement existing measures such as pupillometry.

6.4 Overall Limitations of the Experimental Work Presented in the Thesis

6.4.1 Correlational design

A correlational design was used in Chapters 3 and 5 to assess the contribution of cognition and eye gaze respectively, during recognition of, and perceptual adaptation to, unfamiliar speech. The advantage of such a design is that it highlights individual variation in response to different conditions. However, the main disadvantage is that causation cannot be inferred. In Chapter 3, this was not problematic since the aim was

to identify the cognitive abilities that are involved in perceptual adaptation, not that specifically cause it (that is, the aim was not to establish whether an individual's ability to recognise accented speech develops because they have a good vocabulary, or vice versa). For Study 5, however, conclusions as to the role of eye gaze in relation to audiovisual speech recognition are limited, as it unknown whether eye gaze strategies result from an individual's ability to recognise audiovisual speech, or vice versa. This will need to be established in future research.

A second problem with correlational designs is that of inter-correlations between variables. This is specifically problematic in Chapter 3, in which the relative contribution of several variables was assessed. Although all assumptions for carrying out a multiple regression analysis were checked, including significant collinearity, the overlap between the working memory test and vocabulary test (and the correlation between them) could account for the mediation effect that was observed. Further experimentation, using a different type of working memory test for example, is needed to verify this finding.

Finally, to accurately establish the contribution of any predictor variable in a correlational design, all possible variables need to be included into the model. Practically, this is rarely possible in a single psychological study, not least because a very large sample size would also be required. However, it should be noted that only a small proportion of the variation in recognition of unfamiliar speech, and perceptual adaptation to unfamiliar speech, has been explained by the present studies. This indicates that further predictors need to be identified to fully explain individual differences in perceptual plasticity. These could be cognitive or sensory abilities, as well as factors such as motivation (Sygal Amitay, Halliday, Taylor, Sohoglu, & Moore, 2010), concentration, or even health and lifestyle factors (Dawes et al., 2014). Whatever the exact predictors, all possible variables need to be present in a statistical model for their contribution to be accurately established, and to fully explain individual differences in perceptual plasticity.

6.4.2 IEEE sentences

As discussed in the General Methods, the same IEEE sentences were used in all three studies reported here for consistency. These sentences are used widely in

audiological and psychological research, and are often described as being standardised and semantically unpredictable. Nevertheless, there was some variability evident in participant responses, and some items were likely more difficult than others (although this could equally have been due to variation in the speakers pronunciation, especially with the Japanese-accented speech). Since the effects of exposure to multiple sentences were analysed over time, an items analysis to examine item-specific effects in the data was not required. Furthermore, sentences were fully randomised and counterbalanced to prevent any item-specific effects. Nevertheless, the IEEE sentences may not be the most suitable stimuli for studies of speech recognition for several reasons.

Firstly, there is some variation in their predictability; for example, in the sentence "Sunday is the best part of the week", the object "week" could easily be predicted from the subject "Sunday" due to the close semantic association. In comparison, for the sentence "The friendly gang left the drug store", the object "drug store" is not easily predictable from the subject "friendly gang". Secondly, the familiarity of the words included in them varies, and some words may seem slightly old-fashioned to younger listeners; for example, "the wharf" is a more uncommon word than "the tree", while the phrases "to suffer fright" or "drug store" are not commonly used in modern spoken British English. Thirdly, the sentences vary in the complexity of their syntactic structure; for example, they vary from simple subject-verb-object constructions such as "The tiny girl took off her hat", to more complex, and less predictable constructions, such as "The harder he tried the less he got done".

Such variations are not ideal for use in a study that is dependent on each trial being of equal difficulty, as they may introduce confounds relating to item-specific effects. These limitations of the IEEE sentences should be taken into account in future studies considering using them. However, the steps taken in the present studies to counter any such confounds, namely randomisation, counterbalancing and analysis of responses to multiple rather than individual sentences, should have prevented any such confounds from affecting the results.

6.4.3 Background noise

A major limitation to the studies reported in chapters 3 and 4 is the use of background noise to control for ceiling effects in recognition of the accented speech. As

listeners can adapt very rapidly to accented speech after just a few sentences (e.g. Clarke & Garrett, 2004), we added background noise to prevent participants from reaching ceiling levels of accuracy before completing the experiment. In fact, perceptual adaptation in Chapter 3, and in Study 1 of Chapter 4, was measured in terms of participants' tolerance to background noise. Background noise has been used in a similar way in several previous studies of accented speech recognition (Adank & Janse, 2010; Bradlow & Bent, 2007; Janse & Adank, 2012; Yi et al., 2013). Nevertheless, the use of background noise in the present work limits the conclusions that can be drawn from it somewhat. Particularly, effects relating to the accented speech and those relating to the background noise cannot be separated.

In Chapter 3, this problem was addressed by including participants' baseline recognition of standard British English speech in background noise, as a predictor variable in our regression and path models. In this way, an attempt was made to control for participants' ability to process 'unaccented' speech in background noise. In Study 1 of Chapter 4, this was not possible (a covariate is not easily included in a repeated measures analysis), and so the question of whether audiovisual speech cues can benefit recognition of accented speech remains unanswered. Future studies may be able to address this problem by developing a method for analysing perceptual adaptation to accented speech, without the addition of background noise.

6.5 Conclusion

In recent years, there has been an increase in research into how listeners respond to adverse listening conditions related to the speech source, as researchers have begun to recognise that cognitive and behavioural plasticity in such contexts is vital for successful communication. The results reported in this thesis demonstrate how this perceptual plasticity is influenced by different factors relating to the listener and to the listening context, such as the presence of audiovisual cues. Our results support mounting evidence that recognition of, and adaptation to, unfamiliar speech is driven by cognitive processes, in particular executive function and vocabulary skill. They further demonstrate that recognition of unfamiliar speech is improved when audiovisual cues

are available, and that listeners' eye gaze may play a role in the benefits that can be gained from them.

The findings presented here also reveal a clear distinction between measurements of perceptual plasticity; in each study, the specific factors under investigation have differentially affected overall recognition and perceptual adaptation. While recognition of unfamiliar speech was related to individual differences in vocabulary skill, working memory, eye gaze and the speech modality, perceptual adaptation was only influenced by one measure of cognition – inhibitory abilities. These different results suggest that speech recognition and perceptual adaptation are different abilities, driven and affected by different factors relating to the listener and the environment. Although the measurements have mostly been studied independently of one another (although, see Janse & Adank, 2012), a clear distinction between them has not been made. In fact, speech recognition is often taken as the key measurement of how well listeners deal with adverse listening conditions. Whilst it is important to understand listeners' overall ability to deal with adverse conditions, it is also important to consider their capacity to learn, and the factors that could affect this. Future studies which aim to understand the mechanisms of perceptual plasticity in relation to adverse listening conditions, should therefore take both types of response into consideration.

References

Adank, P., Evans, B. G., Stuart-Smith, J., & Scott, S. K. (2009). Comprehension of Familiar and Unfamiliar Native Accents Under Adverse Listening Conditions. *Journal of Experimental Psychology-Human Perception and Performance, 35*(2), 520-529. doi:10.1037/a0013552

Adank, P., & Janse, E. (2010). Comprehension of a Novel Accent by Young and Older Listeners. *Psychology and Aging, 25*(3), 736-740. doi:10.1037/a0020054

Ahissar, M., & Hochstein, S. (2004). The reverse hierarchy theory of visual perceptual learning. *Trends in Cognitive Sciences, 8*(10), 457-464. doi:10.1016/j.tics.2004.08.011

Akeroyd, M. A. (2008). Are individual differences in speech reception related to individual differences in cognitive ability? A survey of twenty experimental studies with normal and hearing-impaired adults. *International Journal of Audiology, 47*, S53-S71. doi:Pii 90555050910.1080/14992020802301142

Amitay, S. (2009). Forward and Reverse Hierarchies in Auditory Perceptual Learning. *Learning and Perception, 1*(1), 59-68.

Amitay, S., Halliday, L., Taylor, J., Sohoglu, E., & Moore, D. R. (2010). Motivation and Intelligence Drive Auditory Perceptual Learning. *Plos One, 5*(3). doi:10.1371/journal.pone.0009816

Arnold, P., & Hill, F. (2001). Bisensory augmentation: A speechreading advantage when speech is clearly audible and intact. *British Journal of Psychology, 92*, 339-355.

Baese-Berk, M. M., Bradlow, A. R., & Wright, B. A. (2013). Accent-independent adaptation to foreign accented speech. *Journal of the Acoustical Society of America, 133*(3), EL174-EL180. doi:10.1121/1.4789864

Banks, B., Gowen, E., Munro, K., & Adank, P. (2015). Cognitive predictors of perceptual adaptation to accented speech. *Journal of the Acoustical Society of America, 137*(4), 2015-2024. doi:10.1121/1.4916265

Bernstein, L. E., Auer, E. T., Jr., Eberhardt, S. P., & Jiang, J. (2013). Auditory perceptual learning for speech perception can be enhanced by audiovisual training. *Frontiers in Neuroscience, 7*. doi:10.3389/fnins.2013.00034

Besser, J., Koelewijn, T., Zekveld, A. A., Kramer, S. E., & Festen, J. M. (2013). How Linguistic Closure and Verbal Working Memory Relate to Speech Recognition in Noise-A Review. *Trends in Amplification, 17*(2), 75-93. doi:10.1177/1084713813495459

Birmingham, E., & Kingstone, A. (2009). Human Social Attention A New Look at Past, Present, and Future Investigations. *Year in Cognitive Neuroscience 2009, 1156*, 118-140. doi:10.1111/j.1749-6632.2009.04468.x

Boersma, P., & Weenink, D. (2012). Praat: doing phonetics by computer (Version 5.3.05).

Boothroyd, A. (2010). Adapting to Changed Hearing: The Potential Role of Formal Training. *Journal of the American Academy of Audiology, 21*(9), 601-611. doi:10.3766/jaaa.21.9.6

Borovsky, A., Elmana, J. L., & Fernald, A. (2012). Knowing a lot for one's age: Vocabulary skill and not age is associated with anticipatory incremental sentence interpretation in children and adults. *Journal of Experimental Child Psychology, 112*(4), 417-436. doi:10.1016/j.jecp.2012.01.005

Borrie, S. A., McAuliffe, M. J., & Liss, J. M. (2012). Perceptual Learning of Dysarthric Speech: A Review of Experimental Studies. *Journal of Speech Language and Hearing Research, 55*(1), 290-305. doi:10.1044/1092-4388(2011/10-0349)

Bradlow, A. R., Akahane-Yamada, R., Pisoni, D. B., & Tohkura, Y. (1999). Training Japanese listeners to identify English vertical bar r vertical bar and vertical bar l vertical bar: Long-term retention of learning in perception and production. *Perception & Psychophysics, 61*(5), 977-985. doi:10.3758/bf03206911

Bradlow, A. R., & Bent, T. (2008). Perceptual adaptation to non-native speech. *Cognition, 106*(2), 707-729. doi:10.1016/j.cognition.2007.04.005

Bradlow, A. R., Pisoni, D. B., AkahaneYamada, R., & Tohkura, Y. (1997). Training Japanese listeners to identify English vertical bar r vertical bar and vertical bar l vertical bar .4. Some effects of perceptual learning on speech product. *Journal of the Acoustical Society of America, 101*(4), 2299-2310. doi:10.1121/1.418276

Bruce, C., To, C.-T., & Newton, C. (2012). Accent on communication: the impact of regional and foreign accent on comprehension in adults with aphasia. *Disability and Rehabilitation, 34*(12). doi:10.3109/09638288.2011.631680

Buchan, J. N., Pare, M., & Munhall, K. G. (2007). Spatial statistics of gaze fixations during dynamic face processing. *Social Neuroscience, 2*(1), 1-13. doi:10.1080/17470910601043644

Buchan, J. N., Pare, M., & Munhall, K. G. (2008). The effect of varying talker identity and listening conditions on gaze behavior during audiovisual speech perception. *Brain Research, 1242*, 162-171. doi:10.1016/j.brainres.2008.06.083

Christianson, S. A., Loftus, E. F., Hoffman, H., & Loftus, G. R. (1991). Eye fixations and memory for emotional events. *Journal of Experimental Psychology-Learning Memory and Cognition, 17*(4), 693-701. doi:10.1037//0278-7393.17.4.693

Clarke, C. M., & Garrett, M. F. (2004). Rapid adaptation to foreign-accented English. *Journal of the Acoustical Society of America, 116*(6), 3647-3658. doi:10.1121/1.1815131

Cristia, A., Seidl, A., Vaughn, C., Schmale, R., Bradlow, A., & Floccia, C. (2012). Linguistic processing of accented speech across the lifespan. *Frontiers in psychology, 3*, 479-479. doi:10.3389/fpsyg.2012.00479

Daneman, M., & Carpenter, P. A. (1980). Individual-differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior, 19*(4), 450-466. doi:10.1016/s0022-5371(80)90312-6

Davis, M. H., & Johnsrude, I. S. (2003). Hierarchical processing in spoken language comprehension. *Journal of Neuroscience, 23*(8), 3423-3431.

Davis, M. H., & Johnsrude, I. S. (2007). Hearing speech sounds: Top-down influences on the interface between audition and speech perception. *Hearing Research, 229*(1-2), 132-147. doi:10.1016/j.heares.2007.01.014

Davis, M. H., Johnsrude, I. S., Hervais-Adelman, A., Taylor, K., & McGettigan, C. (2005). Lexical information drives; Perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences. *Journal of Experimental Psychology-General, 134*(2), 222-241. doi:10.1037/0096-3445.134.2.222

Dawes, P., Cruickshanks, K. J., Moore, D. R., Edmondson-Jones, M., McCormack, A., Fortnum, H., & Munro, K. J. (2014). Cigarette Smoking, Passive Smoking, Alcohol Consumption, and Hearing Loss. *Jaro-Journal of the Association for Research in Otolaryngology, 15*(4), 663-674. doi:10.1007/s10162-014-0461-0

De Rivecourt, M., Kuperus, M. N., Post, W. J., & Mulder, L. J. M. (2008). Cardiovascular and eye activity measures as indices for momentary changes in mental effort during simulated flight. *Ergonomics, 51*(9), 1295-1319. doi:10.1080/00140130802120267

Deubel, H., & Schneider, W. X. (1996). Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision Research, 36*(12), 1827-1837. doi:10.1016/0042-6989(95)00294-4

Dupoux, E., & Green, K. (1997). Perceptual adjustment to highly compressed speech: Effects of talker and rate changes. *Journal of Experimental Psychology-Human Perception and Performance, 23*(3), 914-927. doi:10.1037//0096-1523.23.3.914

Eisner, F., & McQueen, J. M. (2005). The specificity of perceptual learning in speech processing. *Perception & Psychophysics, 67*(2), 224-238. doi:10.3758/bf03206487

Eisner, F., & McQueen, J. M. (2006). Perceptual learning in speech: Stability over time (L). *Journal of the Acoustical Society of America, 119*(4), 1950-1953. doi:10.1121/1.2178721

Ellis, R. J., & Munro, K. J. (2013). Does cognitive function predict frequency compressed speech recognition in listeners with normal hearing and normal cognition? *International Journal of Audiology, 52*(1), 14-22. doi:10.3109/14992027.2012.721013

Erb, J., Henry, M. J., Eisner, F., & Obleser, J. (2012). Auditory skills and brain morphology predict individual differences in adaptation to degraded speech. *Neuropsychologia, 50*(9), 2154-2164. doi:10.1016/j.neuropsychologia.2012.05.013

Erber, N. P. (1975). Auditory-visual perception of speech. *Journal of Speech and Hearing Disorders, 40*(4), 481-492.

Everdell, I. T., Marsh, H., Yurick, M. D., Munhall, K. G., & Pare, M. (2007). Gaze behaviour in audiovisual speech perception: Asymmetrical distribution of face-directed fixations. *Perception, 36*(10), 1535-1545. doi:10.1068/p5852

Faulkner, A., Rosen, S., & Smith, C. (2000). Effects of the salience of pitch and periodicity information on the intelligibility of four-channel vocoded speech: Implications for cochlear implants. *Journal of the Acoustical Society of America, 108*(4), 1877-1887.

Floccia, C., Goslin, J., Girard, F., & Konopczynski, G. (2006). Does a regional accent perturb speech processing? *Journal of Experimental Psychology-Human Perception and Performance, 32*(5), 1276-1293. doi:10.1037/0096-1523.32.5.1276

Giraud, A. L., Kell, C., Thierfelder, C., Sterzer, P., Russ, M. O., Preibisch, C., & Kleinschmidt, A. (2004). Contributions of sensory input, auditory search and verbal comprehension to cortical activity during speech processing. *Cerebral Cortex, 14*(3), 247-255. doi:10.1093/cercor/bhg124

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying fixed effects analyses of variance and covariance. *Review of Educational Research, 42*(3), 237-288. doi:10.3102/00346543042003237

Goldinger, S. D. (1996). Words and voices: Episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology-Learning Memory and Cognition, 22*(5), 1166-1183. doi:10.1037/0278-7393.22.5.1166

Goldstone, R. L. (1998). Perceptual learning. *Annual Review of Psychology, 49*, 585-612.

Golomb, J. D., Peelle, J. E., & Wingfield, A. (2007). Effects of stimulus variability and adult aging on adaptation to time-compressed speech. *Journal of the Acoustical Society of America, 121*(3), 1701-1708. doi:10.1121/1.2436635

Gordon-Salant, S., & Fitzgibbons, P. J. (1993). Temporal factors and speech recognition performance in young and elderly listeners. *Journal of Speech and Hearing Research, 36*(6), 1276-1285.

Gordon-Salant, S., & Fitzgibbons, P. J. (2001). Sources of age-related recognition difficulty for time-compressed speech. *Journal of Speech Language and Hearing Research, 44*(4), 709-719. doi:10.1044/1092-4388(2001/056)

Gordon-Salant, S., Yeni-Komshian, G. H., & Fitzgibbons, P. J. (2010). Short-term adaptation to accented English by younger and older adults. *Journal of the Acoustical Society of America, 128*(4), E200-E204. doi:10.1121/1.3486199

Gordon-Salant, S., Yeni-Komshian, G. H., Fitzgibbons, P. J., Cohen, J. I., & Waldroup, C. (2013). Recognition of accented and unaccented speech in different maskers by younger and older listeners. *Journal of the Acoustical Society of America, 134*(1), 618-627. doi:10.1121/1.4807817

Grant, K. W., & Seitz, P. F. (2000). The use of visible speech cues for improving auditory detection of spoken sentences. *Journal of the Acoustical Society of America, 108*(3), 1197-1208. doi:10.1121/1.1288668

Grant, K. W., Walden, B. E., & Seitz, P. F. (1998). Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration. *Journal of the Acoustical Society of America, 103*(5), 2677-2690. doi:10.1121/1.422788

Green, T., Rosen, S., Faulkner, A., & Paterson, R. (2013). Adaptation to spectrally-rotated speech. *Journal of the Acoustical Society of America, 134*(2), 1369-1377. doi:10.1121/1.4812759

Greenspan, S. L., Nusbaum, H. C., & Pisoni, D. B. (1988). Perceptual-learning of synthetic speech produced by rule. *Journal of Experimental Psychology-Learning Memory and Cognition, 14*(3), 421-433. doi:10.1037/0278-7393.14.3.421

Greenwood, D. D. (1990). A cochlear frequency-position function for several species - 29 years later. *Journal of the Acoustical Society of America, 87*(6), 2592-2605. doi:10.1121/1.399052

Hailstone, J. C., Ridgway, G. R., Bartlett, J. W., Goll, J. C., Crutch, S. J., & Warren, J. D. (2012). Accent processing in dementia. *Neuropsychologia, 50*(9). doi:10.1016/j.neuropsychologia.2012.05.027

Halliday, L. F., Moore, D. R., Taylor, J. L., & Amitay, S. (2011). Dimension-specific attention directs learning and listening on auditory training tasks. *Attention Perception & Psychophysics, 73*(5), 1329-1335. doi:10.3758/s13414-011-0148-0

Harwell, M. R., Rubinstein, E. N., Hayes, W. S., & Olds, C. C. (1992). Summarizing monte-carlo results in methodological research - the 1-factor and 2-factor fixed effects anova cases. *Journal of Educational Statistics, 17*(4), 315-339. doi:10.3102/10769986017004315

Hawkins, S. (2003). Roles and representations of systematic fine phonetic detail in speech understanding. *Journal of Phonetics, 31*(3-4), 373-405. doi:10.1016/j.wocn.2003.09.006

Hawley, M. L., Litovsky, R. Y., & Culling, J. F. (2004). The benefit of binaural hearing in a cocktail party: Effect of location and type of interferer. *Journal of the Acoustical Society of America, 115*(2), 833-843. doi:10.1121/1.1639908

Hazan, V., Kim, J., & Chen, Y. (2010). Audiovisual perception in adverse conditions: Language, speaker and listener effects. *Speech Communication, 52*(11-12), 996-1009. doi:10.1016/j.specom.2010.05.003

Hazan, V., Sennema, A., Iba, M., & Faulkner, A. (2005). Effect of audiovisual perceptual training on the perception and production of consonants by Japanese learners of English. *Speech Communication, 47*(3), 360-378. doi:10.1016/j.specom.2005.04.007

Henderson, J. M., Malcolm, G. L., & Schandl, C. (2009). Searching in the dark: Cognitive relevance drives attention in real-world scenes. *Psychonomic Bulletin & Review, 16*(5), 850-856. doi:10.3758/pbr.16.5.850

Hervais-Adelman, A., Davis, M. H., Johnsrude, I. S., & Carlyon, R. P. (2008). Perceptual learning of noise vocoded words: Effects of feedback and lexicality. *Journal of Experimental Psychology-Human Perception and Performance, 34*(2), 460-474. doi:10.1037/0096-1523.34.2.460

Huettig, F., & Janse, E. (2015). Individual differences in working memory and processing speed predict anticipatory spoken language processing in the visual world. *Language, Cognition and Neuroscience*. doi:10.1080/23273798.2015.1047459

Huettig, F., Rommers, J., & Meyer, A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta Psychologica, 137*(2), 151-171. doi:10.1016/j.actpsy.2010.11.003

Huyck, J. J., & Johnsrude, I. S. (2012). Rapid perceptual learning of noise-vocoded speech requires attention. *Journal of the Acoustical Society of America, 131*(3), EL236-EL242. doi:10.1121/1.3685511

IEEE. (1969). Ieee recommended practice for speech quality measurements. *IEEE Transactions on Audio and Electroacoustics, AU17*(3), 225-&.

Janse, E. (2009). Processing of fast speech by elderly listeners. *Journal of the Acoustical Society of America, 125*(4), 2361-2373. doi:10.1121/1.3082117

Janse, E., & Adank, P. (2012). Predicting foreign-accent adaptation in older adults. *Quarterly Journal of Experimental Psychology, 65*(8), 1563-1585. doi:10.1080/17470218.2012.658822

Javal, L. E. (1878). Essai sur la physiologie de la lecture. *Annales d'Oculistique, 79*, 97-117.

Kamphaus, R. W. (2005). *Clinical Assessment of Child and Adolescent Intelligence* (2nd ed.). New York: Springer Science.

Kawase, S., Hannah, B., & Wang, Y. (2014). The influence of visual speech information on the intelligibility of English consonants produced by non-native speakers. *Journal of the Acoustical Society of America, 136*(3), 1352-1362. doi:10.1121/1.4892770

Kawase, T., Sakamoto, S., Hori, Y., Maki, A., Suzuki, Y., & Kobayashi, T. (2009). Bimodal audio-visual training enhances auditory adaptation process. *NeuroReport, 20*(14), 1231-1234.

Kennedy, K. M., Rodrigue, K. M., Head, D., Gunning-Dixon, F., & Raz, N. (2009). Neuroanatomical and Cognitive Mediators of Age-Related Differences in Perceptual Priming and Learning. *Neuropsychology, 23*(4), 475-491. doi:10.1037/a0015377

Kowler, E. (2011). Eye movements: The past 25 years. *Vision Research, 51*(13), 1457-1483. doi:10.1016/j.visres.2010.12.014

Kraljic, T., & Samuel, A. G. (2006). Generalization in perceptual learning for speech. *Psychonomic Bulletin & Review, 13*(2), 262-268. doi:10.3758/bf03193841

Kraljic, T., & Samuel, A. G. (2011). Perceptual learning evidence for contextually-specific representations. *Cognition, 121*(3), 459-465. doi:10.1016/j.cognition.2011.08.015

Kricos, P. B., & Lesner, S. A. (1982). Differences in visual intelligibility across talkers. *Volta Review, 84*(4), 219-225.

Kricos, P. B., & Lesner, S. A. (1985). Effect of talker differences on the speechreading of hearing-impaired teenagers. *Volta Review, 87*(1), 5-14.

Langton, S. R. H., Watt, R. J., & Bruce, V. (2000). Do the eyes have it? Cues to the direction of social attention. *Trends in Cognitive Sciences, 4*(2), 50-59. doi:10.1016/s1364-6613(99)01436-9

Lansing, C. R., & McConkie, G. W. (1994). A new method for speechreading research: Tracking observers' eye movements. *Journal of the Academy of Rehabilitative Audiology, 27*, 25-43.

Lansing, C. R., & McConkie, G. W. (1999). Attention to facial regions in segmental and prosodic visual speech perception tasks. *Journal of Speech Language and Hearing Research, 42*(3), 526-539.

Lansing, C. R., & McConkie, G. W. (2003). Word identification and eye fixation locations in visual and visual-plus-auditory presentations of spoken sentences. *Perception & Psychophysics, 65*(4), 536-552. doi:10.3758/bf03194581

Lively, S. E., Pisoni, D. B., Yamada, R. A., Tohkura, Y., & Yamada, T. (1994). Training japanese listeners to identify english vertical-bar-r-vertical-bar and vertical-bar-l-

vertical-bar .3. Long-term retention of new phonetic categories. *Journal of the Acoustical Society of America, 96*(4), 2076-2087. doi:10.1121/1.410149

Liversedge, S. P., & Findlay, J. M. (2000). Saccadic eye movements and cognition. *Trends in Cognitive Sciences, 4*(1), 6-14. doi:10.1016/s1364-6613(99)01418-7

Lix, L. M., Keselman, J. C., & Keselman, H. J. (1996). Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance F test. *Review of Educational Research, 66*(4), 579-619. doi:10.3102/00346543066004579

Loebach, J. L., Pisoni, D. B., & Svirsky, M. A. (2010). Effects of Semantic Context and Feedback on Perceptual Learning of Speech Processed Through an Acoustic Simulation of a Cochlear Implant. *Journal of Experimental Psychology-Human Perception and Performance, 36*(1), 224-234. doi:10.1037/a0017609

Loftus, G. R. (1981). Tachistoscopic simulations of eye fixations on pictures. *Journal of Experimental Psychology-Human Learning and Memory, 7*(5), 369-376.

Loftus, G. R. (1983). Eye fixations on text and scenes. In K. Rayner (Ed.), *Eye movements in reading: perceptual and language processes* (pp. 359-376). New York: Academic Press.

Loizou, P. C., Mani, A., & Dorman, M. F. (2003). Dichotic speech recognition in noise using reduced spectral cues. *Journal of the Acoustical Society of America, 114*(1), 475-483. doi:10.1121/1.1582861

Luce, P. A., & Pisoni, D. B. (1998). Recognizing spoken words: The neighborhood activation model. *Ear and Hearing, 19*(1), 1-36. doi:10.1097/00003446-199802000-00001

Lyxell, B., & Holmberg, I. (2000). Visual speechreading and cognitive performance in hearing-impaired and normal hearing children (11-14 years). *British Journal of Educational Psychology, 70*, 505-518. doi:10.1348/000709900158272

Lyxell, B., & Ronnberg, J. (1989). Information-processing skill and speech-reading. *British Journal of Audiology, 23*(4), 339-348. doi:10.3109/03005368909076523

Macleod, A., & Summerfield, Q. (1987). Quantifying the contribution of vision to speech perception in noise. *British Journal of Audiology, 21*(2), 131-142. doi:10.3109/03005368709077786

Macleod, C. M. (1991). Half a century of research on the stroop effect - an integrative review. *Psychological Bulletin, 109*(2), 163-203. doi:10.1037//0033-2909.109.2.163

Matin, E. (1974). Saccadic suppression - review and an analysis. *Psychological Bulletin, 81*(12), 899-917. doi:10.1037/h0037368

Mattys, S. L., Davis, M. H., Bradlow, A. R., & Scott, S. K. (2012). Speech recognition in adverse conditions: A review. *Language and Cognitive Processes, 27*(7-8), 953-978. doi:10.1080/01690965.2012.705006

Mattys, S. L., & Scharenborg, O. (2014). Phoneme Categorization and Discrimination in Younger and Older Adults: A Comparative Analysis of Perceptual, Lexical, and Attentional Factors. *Psychology and Aging, 29*(1), 150-162. doi:10.1037/a0035387

Mattys, S. L., Seymour, F., Attwood, A. S., & Munafo, M. R. (2013). Effects of Acute Anxiety Induction on Speech Perception: Are Anxious Listeners Distracted Listeners? *Psychological Science, 24*(8), 1606-1608. doi:10.1177/0956797612474323

May, J. G., Kennedy, R. S., Williams, M. C., Dunlap, W. P., & Brannan, J. R. (1990). Eye-movement indexes of mental workload. *Acta Psychologica, 75*(1), 75-89. doi:10.1016/0001-6918(90)90067-p

Maye, J., Aslin, R. N., & Tanenhaus, M. K. (2008). The Weckud Wetch of the Wast: Lexical adaptation to a novel accent. *Cognitive Science, 32*(3), 543-562. doi:10.1080/03640210802035357

Mcgarr, N. S. (1983). The Intelligibility of Deaf Speech to Experienced and Inexperienced Listeners. *Journal of Speech and Hearing Research, 26*(3), 451-458.

McGurk, H., & Macdonald, J. (1976). HEARING LIPS AND SEEING VOICES. *Nature, 264*(5588), 746-748. doi:10.1038/264746a0

McQueen, J. M., Norris, D., & Cutler, A. (2006). The dynamic nature of speech perception. *Language and Speech, 49*, 101-112.

Miller, G. A. (1947). The masking of speech. *Psychological Bulletin, 44*(2), 105-129. doi:10.1037/h0055960

Miyake, A., Friedman, N. P., Emerson, M. J., Witzki, A. H., Howerter, A., & Wager, T. D. (2000). The unity and diversity of executive functions and their contributions to complex "frontal lobe" tasks: A latent variable analysis. *Cognitive Psychology, 41*(1), 49-100. doi:10.1006/cogp.1999.0734

Moore, B. C. J. (1998). *Cochlear Hearing Loss: Physiological, Psychological and Technical Issues* (2nd ed.). Chichester, UK: John Wiley & Sons. Ltd.

Neger, T. M., Rietveld, T., & Janse, E. (2014). Relationship between perceptual learning in speech and statistical learning in younger and older adults. *Frontiers in Human Neuroscience, 8*. doi:10.3389/fnhum.2014.00628

Newton, C., Burns, R., & Bruce, C. (2013). Accent identification by adults with aphasia. *Clinical Linguistics & Phonetics, 27*(4), 287-298. doi:10.3109/02699206.2012.753111

Norris, D., McQueen, J. M., & Cutler, A. (2003). Perceptual learning in speech. *Cognitive Psychology, 47*(2), 204-238. doi:Doi 10.1016/S0010-0285(03)00006-9

Pallier, C., Sebastian-Galles, N., Dupoux, E., Christophe, A., & Mehler, J. (1998). Perceptual adjustment to time-compressed speech: A cross-linguistic study. *Memory & Cognition, 26*(4), 844-851.

Peelle, J. E., & Wingfield, A. (2005). Dissociations in perceptual learning revealed by adult age differences in adaptation to time-compressed speech. *Journal of Experimental Psychology-Human Perception and Performance, 31*(6), 1315-1330. doi:10.1037/0096-1523.31.6.1315

Pichora-Fuller, M. K. (2003). Cognitive aging and auditory information processing. *International Journal of Audiology, 42*, S26-S32.

Pilling, M., & Thomas, S. (2011). Audiovisual Cues and Perceptual Learning of Spectrally Distorted Speech. *Language and Speech, 54*, 487-497. doi:10.1177/0023830911404958

Pinet, M., Iverson, P., & Evans, B. (2011). *Perceptual Adaptation for L1 and L2 Accents in Noise by Monolingual British English Listeners.* Paper presented at the Proceedings of the International Congress of Phonetic Sciences., Hong Kong.

Pisoni, D. B. (1997). Some Thoughts on "Normalization" in Speech Perception. In K. a. M. Johnson, J.W. (Ed.), *Talker Variability in Speech Processing* (pp. 9-32). San Diego: Academic Press.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*(3), 372-422. doi:10.1037/0033-2909.124.3.372

Rayner, K., & Reichle, E. D. (2010). Models of the reading process. *Wiley Interdisciplinary Reviews-Cognitive Science, 1*(6), 787-799. doi:10.1002/wcs.68

Rizzolatti, G., Riggio, L., & Sheliga, B. M. (1994). SPACE AND SELECTIVE ATTENTION. *Attention and Performance Xv: Conscious and Nonconscious Information Processing, 15*, 231-265.

Ronnberg, J., Lyxell, B., Arlinger, S., & Kinnefors, C. (1989). Visual evoked-potentials - relation to adult speechreading and cognitive function. *Journal of Speech and Hearing Research, 32*(4), 725-735.

Ronnberg, J., Rudner, M., Foo, C., & Lunner, T. (2008). Cognition counts: A working memory system for ease of language understanding (ELU). *International Journal of Audiology, 47*, S99-S105. doi:10.1080/14992020802301167

Rosen, S., Souza, P., Ekelund, C., & Majeed, A. A. (2013). Listening to speech in a background of other talkers: Effects of talker number and noise vocoding. *Journal of the Acoustical Society of America, 133*(4), 2431-2443. doi:10.1121/1.4794379

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2007). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environment. *Cerebral Cortex, 17*(5), 1147-1153. doi:10.1093/cercor/bhl024

Samuel, A. G., & Kraljic, T. (2009). Perceptual learning for speech. *Attention Perception & Psychophysics, 71*(6), 1207-1218. doi:10.3758/app.71.6.1207

Schaie, K. W., Willis, S. L., & Ohanlon, A. M. (1994). Perceived intellectual-performance change over 7 years. *Journals of Gerontology, 49*(3), P108-P118.

Scharenborg, O., Weber, A., & Janse, E. (2015). The role of attentional abilities in lexically guided perceptual learning by older listeners. *Attention Perception & Psychophysics, 77*(2), 493-507. doi:10.3758/s13414-014-0792-2

Schmale, R., Cristia, A., & Seidl, A. (2012). Toddlers recognize words in an unfamiliar accent after brief exposure. *Developmental Science, 15*(6), 732-738. doi:10.1111/j.1467-7687.2012.01175.x

Schneider, B. A., Daneman, M., & Murphy, D. R. (2005). Speech comprehension difficulties in older adults: Cognitive slowing or age-related changes in hearing? *Psychology and Aging, 20*(2), 261-271. doi:10.1037/0882-7974.20.2.261

Schulz, C. M., Schneider, E., Fritz, L., Vockeroth, J., Hapfelmeier, A., Brandt, T., . . . Schneider, G. (2011). Visual attention of anaesthetists during simulated critical incidents. *British Journal of Anaesthesia, 106*(6), 807-813. doi:10.1093/bja/aer087

Schwab, E. C., Nusbaum, H. C., & Pisoni, D. B. (1985). Some effects of training on the perception of synthetic speech. *Human Factors, 27*(4), 395-408.

Scott, S. K., Blank, C. C., Rosen, S., & Wise, R. J. S. (2000). Identification of a pathway for intelligible speech in the left temporal lobe. *Brain, 123*, 2400-2406. doi:10.1093/brain/123.12.2400

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech Recognition with Primarily Temporal Cues. *Science, 270*(5234), 303-304.

Singer, T., Verhaeghen, P., Ghisletta, P., Lindenberger, U., & Baltes, P. B. (2003). The fate of cognition in very old age: Six-year longitudinal findings in the Berlin Aging Study (BASE). *Psychology and Aging, 18*(2), 318-331. doi:10.1037/0882-7974.18.2.318

Sommers, M. S., Tye-Murray, N., & Spehar, B. (2005). Auditory-visual speech perception and auditory-visual enhancement in normal-hearing younger and older adults. *Ear and Hearing, 26*(3), 263-275. doi:10.1097/00003446-200506000-00003

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), 410-424. doi:10.3758/bf03205560

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18*, 643-662. doi:10.1037/0096-3445.121.1.15

Sumby, W., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America, 26*(2), 212-215.

Summerfield, Q. (1987). Some preliminaries to a comprehensive account of audio-visual speech perception. In B. Dodd & R. Campbell (Eds.), *Hearing by Eye: The Psychology of Lip-Reading*. Hillsdale, NJ: Lawrence Erlbaum.

Swerts, M., & Krahmer, E. (2008). Facial expression and prosodic prominence: Effects of modality and facial area. *Journal of Phonetics, 36*(2), 219-238. doi:10.1016/j.wocn.2007.05.001

Treisman, A. M., & Gelade, G. (1980). Feature-integration theory of attention. *Cognitive Psychology, 12*(1), 97-136. doi:10.1016/0010-0285(80)90005-5

Tye-Murray, N., Sommers, M. S., & Spehar, B. (2007). Audiovisual integration and Lipreading abilities of older adults with normal and impaired hearing. *Ear and Hearing, 28*(5), 656-668. doi:10.1097/AUD.0b013e31812f7185

Tyler, R. S., Summerfield, Q., Wood, E. J., & Fernandes, M. A. (1982). Psychoacoustic and phonetic temporal processing in normal and hearing-impaired listeners. *Journal of the Acoustical Society of America, 72*(3), 740-752. doi:10.1121/1.388254

Vatikiotis-Bateson, E., Eigsti, I. M., Yano, S., & Munhall, K. G. (1998). Eye movement of perceivers during audiovisual speech perception. *Perception & Psychophysics, 60*(6), 926-940. doi:10.3758/bf03211929

Vo, M. L. H., Smith, T. J., Mital, P. K., & Henderson, J. M. (2012). Do the eyes really have it? Dynamic allocation of attention when viewing moving faces. *Journal of Vision, 12*(13). doi:10.1167/12.13.3

Wayne, R. V., & Johnsrude, I. S. (2012). The Role of Visual Speech Information in Supporting Perceptual Learning of Degraded Speech. *Journal of Experimental Psychology-Applied, 18*(4), 419-435. doi:10.1037/a0031042

Wechsler, D. (1958). *The Measurement and Appraisal of Adult Intelligence* (4th ed.). Baltimore, MD: The Williams and Wilkins Company.

Wechsler, D. (1999). *Wechsler Abbreviated Scale of Intelligence*. San Antonio, TX: Pearson.

White, K. S., & Aslin, R. N. (2011). Adaptation to novel accents by toddlers. *Developmental Science, 14*(2), 372-384. doi:10.1111/j.1467-7687.2010.00986.x

Wild, C. J., Yusuf, A., Wilson, D. E., Peelle, J. E., Davis, M. H., & Johnsrude, I. S. (2012). Effortful Listening: The Processing of Degraded Speech Depends Critically on Attention. *Journal of Neuroscience, 32*(40), 14010-14021. doi:10.1523/jneurosci.1528-12.2012

Wingfield, A., McCoy, S. L., Peelle, J. E., Tun, P. A., & Cox, L. C. (2006). Effects of adult aging and hearing loss on comprehension of rapid speech varying in syntactic complexity. *Journal of the American Academy of Audiology, 17*(7), 487-497. doi:10.3766/jaaa.17.7.4

Wolfle, L. M. (2003). The introduction of path analysis to the social sciences, and some emergent themes: an annotated bibiography. *Structural Equation Modeling, 10*(1), 1-34.

Yarbus, A. L. (1967). *Eye movements and vision*. New York: Plenum Press.

Yi, H.-G., Phelps, J. E. B., Smiljanic, R., & Chandrasekaran, B. (2013). Reduced efficiency of audiovisual integration for nonnative speech. *Journal of the Acoustical Society of America, 134*(5), EL387-EL393. doi:10.1121/1.4822320

Zekveld, A. A., Kramer, S. E., & Festen, J. M. (2010). Pupil Response as an Indication of Effortful Listening: The Influence of Sentence Intelligibility. *Ear and Hearing, 31*(4), 480-490. doi:10.1097/AUD.0b013e3181d4f251

Zekveld, A. A., Rudner, M., Johnsrude, I. S., & Ronnberg, J. (2013). The effects of working memory capacity and semantic cues on the intelligibility of speech in noise. *Journal of the Acoustical Society of America, 134*(3), 2225-2234. doi:10.1121/1.4817926

Appendix A

Non-parametric Analyses for Chapter 5

The following non-parametric analyses were carried out in addition to those reported for the eye-tracking data in Chapter 5. Each section corresponds to a section of the results in Chapter 5. Skewed variables are first stated for each section, and analyses are then reported in the same format and order as in Chapter 5. Non-parametric analyses have only been carried out where one or more of the variables were significantly skewed; that is, non-parametric analyses have not been carried out where all variables were normally distributed. Recognition accuracy data, and fixation duration data for good and poorer performers, were normally distributed and so were therefore not included in this appendix.

Wilcoxon signed-rank tests (equivalent to a paired-samples t-test) were used to compare measurements of eye gaze between: 1) IAs per group; and 2) testing blocks/time points per IA and per group, with appropriate Bonferroni corrections. Friedman's ANOVA was carried out per group and per IA to test for effects of testing block/time point. Correlations were carried out using Spearman's rho.

Part B. When do perceivers of audiovisual noise-vocoded speech use visual speech cues?

Eye gaze during recognition of individual sentences.

In the audiovisual group, percent fixations on the eyes and mouth, and fixation duration on the eyes, were significantly skewed variables. In the audio-only group, fixation duration for the eyes was also significantly skewed.

Percent fixations. In the first 500ms of sentence presentation, both groups had a greater percentage of fixations on the eyes than the mouth (audiovisual: Mdn_{eyes} = 59%, IQR = 46.0–87.6%; Mdn_{mouth} = 39%, IQR = 9.9–52.3%; z = 3.03, p = 0.002, r = 0.55; audio-only: Mdn_{eyes} = 76%, IQR = 55.3–84.2%; Mdn_{mouth} = 18%, IQR = 9.3–40.3%; z = 4.16, $p < 0.001$, r = 0.77. Overall, the audiovisual group had a greater percentage of fixations on the mouth than the eyes (Mdn_{mouth} = 82%, IQR = 57.1–86.4%; Mdn_{eyes} =

18%, IQR = 12.8–42.1%, $z = 3.75$, $p < 0.001$, $r = 0.68$, whereas the audio-only group had a greater percentage of fixations on the eyes than on the mouth (Mdn_{eyes} = 64%, IQR = 50.0–76.0; Mdn_{mouth} = 29%, IQR = 17.0–42.1%), $z = 3.41$, $p = 0.001$, $r = 0.63$.

In the audiovisual group, there was a significant effect of time for percent fixations on the mouth, $\chi^2(6) = 118.58$, $p < 0.001$, $\phi = 1.99$, and on the eyes, $\chi^2(6) = 106.11$, $p < 0.001$, $\phi = 1.88$. The percentage of fixations on the mouth significantly increased between 500ms (*Mdn* = 39%, IQR = 9.8–52.3%) and 2500ms (*Mdn* = 96%, IQR = 71.1–99.2%), and significantly decreased on the eyes between 500ms (*Mdn* = 59%, IQR = 46.0–87.6%) and 3000ms (*Mdn* = 4%, IQR = 1.49–24.2%), *p*s < 0.003. In the audio-only group, there was a significant effect of time for percent fixations on the mouth, $\chi^2(6) = 22.06$, $p < 0.001$, $\phi = 0.87$, and on the eyes, $\chi^2(6) = 15.34$, $p = 0.018$, $\phi = 0.73$. In the audio-only group, the percentage of fixations on the mouth significantly increased between 500ms (*Mdn* = 18%, IQR = 9.3–40.3%) and 1500ms (*Mdn* = 31%, IQR = 12.8–55.9%), and significantly decreased on the eyes between 500ms (*Mdn* = 76%, IQR = 55.3–84.2%) and 2500ms (*Mdn* = 64%, IQR = 56.5–79.2%), *p*s < 0.003.

Fixation duration. In the audiovisual group, fixations on the mouth were overall significantly longer than fixations on the eyes (Mdn_{mouth} = 1264ms, IQR = 938.5–2058.6ms; Mdn_{eyes} = 373ms, IQR = 262.9–713.2ms), $z = 4.68$, $p < 0.001$, $r = 0.85$, and at every time point (*p*s < 0.007). In the audio-only group, there was no significant difference in the duration of fixations on the eyes and mouth overall (Mdn_{mouth} = 690ms, IQR = 382.4–1049.7ms; Mdn_{eyes} = 669ms, IQR = 428.1–1161.1ms), $z = 0.62$, $p = 0.538$, $r = 0.12$, or at any time point (*p*s > 0.007).

In the audiovisual group, there was a significant effect of time for percent fixations on the mouth, $\chi^2(6) = 136.14$, $p < 0.001$, $\phi = 2.13$, and on the eyes, $\chi^2(6) = 14.96$, $p = 0.021$, $\phi = 0.71$. Fixations on the mouth significantly increased between 500ms (*Mdn* = 774ms, IQR = 465.0–1118.0ms) and 2500ms (*Mdn* = 1600ms, IQR = 1191.7–2400.7ms), while fixations on the eyes significantly increased between 500ms (*Mdn* = 304ms, IQR = 256.7–402.7ms) and 3000ms (*Mdn* = 394ms, IQR = 269.3–790.0ms), *p*s < 0.001.

In the audio-only group, there was a significant effect of time for percent fixations on the mouth, $\chi^2(6) = 72.86$, $p < 0.001$, $\phi = 1.58$, and on the eyes, $\chi^2(6) =$

87.31, $p < 0.001$, $\phi = 1.74$. Fixations on the mouth significantly increased in duration between 500ms (*Mdn* = 513ms, IQR = 288.5–664.5ms) and 2500ms (*Mdn* = 842ms, IQR = 413.5–1243.0ms), while fixations on the eyes increased in duration between 500ms (*Mdn* = 350ms, IQR = 275.5–644.5ms) and 3000ms (*Mdn* = 856ms, IQR = 494.0–1405.0ms), *p*s < 0.001.

Eye gaze during perceptual adaptation to noise-vocoded speech.

In the audiovisual group, data for percentage gaze time on the eyes and mouth, and fixation duration on the eyes were significantly skewed. In the audio-only group, data for percentage gaze time on the eyes and mouth, and percent fixations on the eyes and mouth were significantly skewed.

Where do listeners look? In the audiovisual group, percent fixation time was overall significantly greater on the mouth (*Mdn* = 78%, *IQR* = 58.3–84.0%) than on the eyes (*Mdn* = 22%, *IQR* = 15.8–40.8%), $z = 3.64$, $p < 0.001$, $r = 0.66$, and in every testing block (*p*s < 0.008). In the audio-only group, fixation time was overall significantly greater on the eyes (*Mdn* = 65%, *IQR* = 52.5–76.5%) than on the mouth (*Mdn* = 15%, *IQR* = 6.3–28%), $z = 4.34$, $p < 0.001$, $r = 0.81$, and in every testing block (*p*s < 0.008).

In the audiovisual group, fixations on the mouth (*Mdn* = 938ms, *IQR* = 690.9–1266.5ms) were overall significantly longer than fixations on the eyes (*Mdn* = 310ms, *IQR* = 274.0–417.8ms), $z = 4.76$, $p < 0.001$, $r = 0.87$, and in every testing block (*p*s < 0.008). In the audio-only group, the duration of fixations on the eyes (*Mdn* = 377ms, *IQR* = 349.2–523.7ms) and on the mouth (*Mdn* = 428ms, *IQR* = 331.8–522.2ms) were not significantly different overall, $z = 0.23$, $p = 0.82$, $r = 0.04$, or in any testing block (*p*s > 0.008).

In the audiovisual group, there was no significant difference between the percentage of fixations on the mouth (*Mdn* = 51%, *IQR* = 39.8–65.0%) and eyes (*Mdn* = 47%, *IQR* = 32.0–59.3%) overall, $z = 0.39$, $p = 0.696$, $r = 0.00$, or in any testing block (*p*s > 0.008). In the audio-only group, a higher percentage of fixations fell on the eyes

(*Mdn* = 68%, *IQR* = 53.0–80.5%) than on the mouth (*Mdn* = 13%, *IQR* = 7.5–23.5%), *z* = 4.41, $p < 0.001$, *r* = 0.82, and in every testing block (*p*s < 0.008).

Do patterns of eye gaze change over time?

There was a significant effect of testing block in the audiovisual group for the duration of fixations on the mouth, $\chi^2(5) = 17.13$, $p = 0.004$, $\phi = 0.76$. There was a significant decrease in fixation duration between block 1 (*Mdn* = 974ms, *IQR* = 840.4–1540.7ms) and block 6 (*Mdn* = 817ms, *IQR* = 610.7–1150.7ms), $z = 4.15$, $p < 0.001$, $r = 0.76$.

There was a significant effect of testing block in the audio-only group for percent fixations on the mouth, $\chi^2(5) = 16.88$, $p = 0.005$, $\phi = 0.76$. There was a significant increase in percent fixations between block 1 (*Mdn* = 8%, IQR = 5.3–15.9%) and block 3 (*Mdn* = 15%, IQR = 7.4–22.4%), $z = 3.25$, $p = 0.001$, $r = 0.60$.

Part C. Is eye gaze related to recognition of audiovisual noise-vocoded speech?

Correlational analyses

Data for percent fixation time on the mouth (early and later) were significantly skewed in both groups, and data for percent fixations on the mouth (early and later) were significantly skewed in the audio-only group. Non-parametric correlations for these variables are reported below.

In the audiovisual group, later recognition accuracy was positively correlated with later percent fixation time, $r = 0.36$, $p = 0.049$, CI [0.01, 0.63], indicating that better recognition accuracy was related to more time spent fixating on the speaker's mouth. Early and later recognition accuracy were significantly and positively correlated in the audiovisual group, $r = 0.78$, $p < 0.001$, *CI* = 0.52:0.91, and in the audio-only group, $r = 0.64$, $p < 0.001$, CI [0.30, 0.84]. All eye movement variables (early and later percent fixation time and percent fixations) in both groups, were significantly and positively correlated with one another (see Tables 1.1 and 1.2).

Table 1.1. Audiovisual group

Variable	Mdn	IQR	RA1	RA2	FT1	FT2	F1	F2
RA1 (%)	48	39.1-59.6	-	-	-	-		
RA2 (%)	62	48.3-66.1	0.78**	-	-	-		
FT1 (%)	76	59.8-84.0	0.21	0.29	-	-		
FT2 (%)	79	58.0-83.3	0.26	0.36*	0.92**	-		
F1 (%)	48	39.8-62.0	0.22	0.29	0.92**	0.85**	-	
F2 (%)	52	41.8-65.0	0.23	0.34	0.85**	0.89**	0.95**	-

Table 1.2. Audio-only group

Variable	Mdn	IQR	RA1	RA2	FT1	FT2	F1	F2
RA1 (%)	33	23.1-36.9	-	-	-	-	-	-
RA2 (%)	38	32.2-46.7	0.64**	-	-	-	-	-
FT1 (%)	14	8.2-28.9	-0.13	-0.17	-	-	-	-
FT2 (%)	18	6.8-31.8	-0.10	-0.10	0.80**	-	-	-
F1 (%)	11	7.7-22.7	-0.22	-0.22	0.96**	0.74**	-	-
F2 (%)	16	7.4-2.6	-0.15	-0.20	0.75**	0.92**	0.76**	-

Tables 1.1 and 1.2. Non-parametric correlation matrices for recognition accuracy and eye gaze on the mouth in the audiovisual and audio-only groups. All correlations are Spearman's rho. RA = recognition accuracy; FT = percent fixation time; F = percent fixations; '1' and '2' indicate early and later testing blocks. *$p < 0.05$; **$p < 0.001$.

Appendix B

IEEE sentences and keywords

Native English (baseline) sentences: Chapters 3 and 4	
Full sentence	**Keywords**
Sickness kept him home the third week	sickness kept home week
The wide road shimmered in the hot sun	wide road shimmered sun
Adding fast leads to wrong sums	adding fast leads sums
Wipe the grease off his dirty face	wipe grease dirty face
The meal was cooked before the bell rang	meal cooked bell rang
The small pup gnawed a hole in the sock	pup gnawed hole sock
The fish twisted and turned on the bent hook	fish twisted turned hook
The swan dive was far short of perfect	swan dive short perfect
Hoist the load to your left shoulder	hoist load left shoulder
The box was thrown beside the parked truck	box thrown parked truck
A small creek cut across the field	small creek across field
This is a grand season for hikes on the road	grand season hikes road
Those words were the cue for the actor to leave	words cue actor leave
The ink stain dried on the finished page	ink dried finished page
The walled town was seized without a fight	walled town seized fight
The sky that morning was clear and bright blue	sky morning clear bright
The brown house was on fire to the attic	brown house fire attic
Sunday is the best part of the week	Sunday best part week
The doctor cured him with these pills	doctor cured him pills
The new girl was fired today at noon	new fired today noon
Acid burns holes in wool cloth	acid burns holes cloth
Fairy tales should be fun to write	fairy should fun write
Eight miles of woodland burned to waste	Eight woodland burned waste
The third act was dull and tired the players	third dull tired players
A young child should not suffer fright	young child suffer fright
The club rented the rink for the fifth night	club rented rink night
After the dance they went straight home	after dance went home
The paper box is full of thumb tacks	paper box full tacks
Sell your gift to a buyer at a good gain	sell gift buyer gain
Bring your best compass to the third class	bring best compass class
Novel-accented sentences: Chapters 3 and 4	
Full sentence	**Keywords**

Glue the sheet to the dark blue background	glue sheet dark background
It's easy to tell the depth of a well	easy tell depth well
Rice is often served in round bowls	rice often served bowls
The juice of lemons makes fine punch	juice lemons makes punch
Four hours of steady work faced us	four hours work faced
A large size in stockings is hard to sell	large stockings hard sell
A rod is used to catch pink salmon	rod catch pink salmon
The source of the huge river is the clear spring	source river clear spring
Kick the ball straight and follow through	kick ball straight through
A pot of tea helps to pass the evening	pot tea helps evening
The soft cushion broke the man's fall	soft cushion broke fall
The girl at the stall sold fifty beads	girl stall sold beads
Read verse out loud for pleasure	read verse loud pleasure
Take the winding path to reach the lake	take path reach lake
The wrist was badly strained and hung limp	wrist strained hung limp
The stray cat gave birth to kittens	stray cat birth kittens
The young girl gave no clear response	girl gave clear response
A king ruled the state in the early days	king ruled state days
The ship was torn apart on the sharp reef	ship torn sharp reef
The crooked maze failed to fool the mouse	maze failed fool mouse
The show was a flop from the very start	show flop very start
March the soldiers past the next hill	march soldiers past hill
A cup of sugar makes sweet fudge	cup sugar makes fudge
Place a rosebush near the porch steps	place rosebush porch steps
A steep trail is painful for our feet	steep trail painful feet
We talked of the slide show in the circus	talked side show circus
Use a pencil to write the first draft	pencil write first draft
He ran half way to the hardware store	ran halfway hardware store
The clock struck to mark the third period	clock struck third period
The dune rose from the edge of the water	dune rose edge water
The two met while playing on the sand	two met playing sand
The lease ran out in sixteen weeks	lease out sixteen weeks
A tame squirrel makes a nice pet	tame squirrel nice pet
The pearl was worn in a thin silver ring	pearl worn silver ring
The fruit peel was cut in thick slices	fruit peel cut slices
See the cat glaring at the scared mouse	cat glaring scared mouse
The lawyer tried to lose his case	lawyer tried lose case
Cut the pie into large parts	cut pie large parts
Men strive but seldom get rich	men strive seldom rich
Always close the barn door tight	always close door tight
Bail the boat, to stop it from sinking	bail boat stop sinking

The bill is paid every third week	bill paid third week
Cats and dogs each hate the other	cats dogs hate other
The pipe began to rust while new	pipe began rust new
Thieves who rob friends deserve jail	thieves rob deserve jail
The ripe taste of cheese improves with age	taste cheese improves age
Act on these orders with great speed	act orders great speed
The bark of the pine tree was shiny and dark	bark tree shiny dark
Split the log with a quick, sharp blow	split log quick blow
He ordered peach pie with ice cream	ordered peach pie ice-cream
Weave the carpet on the right hand side	weave carpet right side
We find joy in the simplest things	find joy simplest things
Type out three lists of orders	type three lists orders
The harder he tried the less he got done	harder tried less done
The cup cracked and spilled its contents	cup cracked spilled contents
A cramp is no small danger on a swim	cramp small danger swim
Bring your problems to the wise chief	bring problems wise chief
Write a fond note to the friend you cherish	write note friend cherish
The young kid jumped the rusty gate	kid jumped rusty gate
The just claim got the right verdict	just claim right verdict
Pure bred poodles have curls	pure bred poodles curls
The tree top waved in a graceful way	treetop waved graceful way
The urge to write short stories is rare	urge write stories rare
The pencils have all been used	pencils all been used
The pirates seized the crew of the lost ship	pirates seized crew ship
We tried to replace the coin but failed	tried replace coin failed
The jacket hung on the back of the wide chair	jacket hung back chair
A filing case is now hard to buy	filing case hard buy
The office paint was a dull sad tan	office paint dull tan
He knew the skill of the great young actress	knew skill great actress
A rag will soak up spilt water	rag soak spilt water
A shower of dirt fell from the hot pipes	shower dirt fell pipes
Steam hissed from the broken valve	steam hissed broken valve
The child almost hurt the small dog	child hurt small dog
The sky that morning was clear and bright blue	sky morning clear blue
Sunday is the best part of the week	Sunday best part week
The doctor cured him with these pills	doctor cured these pills
The new girl was fired today at noon	girl fired today noon
Acid burns holes in wool cloth	acid burns holes cloth
Fairy tales should be fun to write	fairy tales fun write
The third act was dull and tired the players	act dull tired players
A young child should not suffer fright	young child suffer fright

We admire and love a good cook	admire love good cook
He carved a head from the round block of marble	carved head round marble
She has a smart way of wearing clothes	smart way wearing clothes
The fruit of a fig tree is apple-shaped	fruit tree apple shaped
Corn cobs can be used to kindle a fire	cobs used kindle fire
Where were they when the noise started	where they noise started
The paper box is full of thumb tacks	box full thumb tacks
Sell your gift to a buyer at a good gain	sell gift buyer gain
The petals fall with the next puff of wind	petals fall puff wind
Bring your best compass to the third class	bring compass third class
The brown house was on fire to the attic	brown house fire attic
The club rented the rink for the fifth night	club rented rink night
After the dance they went straight home	after dance straight home
The hostess taught the new maid to serve	hostess taught maid serve
Grace makes up for lack of beauty	grace makes lack beauty
Nudge gently but wake her now	nudge gently wake now
Bottles hold four kinds of rum	bottles hold kinds rum
The man wore a feather in his felt hat	man wore feather hat
Birth and death mark the limits of life	birth death limits life
The chair looked strong but had no bottom	chair looked strong bottom
Five years he lived with a shaggy dog	five years lived dog
The three story house was built of stone	story house built stone
We like to see clear weather	like see clear weather
The square wooden crate was packed to be shipped	square crate packed shipped
Ripe pears are fit for a queen's table	pears fit queen's table
Hurdle the pit with the aid of a long pole	hurdle pit aid pole
A toad and a frog are hard to tell apart	toad frog hard apart
A round hole was drilled through the thin board	hole drilled thin board
Prod the old mule with a crooked stick	prod mule crooked stick
Dull stories make her laugh	dull stories make laugh
The duke left the park in a silver coach	duke left park coach
Sweet words work better than fierce	words work better fierce
The loss of the cruiser was a blow to the fleet	loss cruiser blow fleet
Plead with the lawyer to drop the lost cause	plead lawyer drop cause
Calves thrive on tender spring grass	calves thrive tender grass
The square peg will settle in the round hole	peg settle round hole
Be sure to set the lamp firmly in the hole	sure lamp firmly hole
Open your book to the first page	open book first page
The long journey home took a year	long journey home year
Small children came to see him	small children see him
There are more than two factors here	more two factors here

We don't get much money but we have fun	get money have fun
Shake hands with this friendly child	shake hands friendly child
If you mumble your speech will be lost	mumble your speech lost
The small red neon lamp went out	small neon lamp out
Breathe deep and smell the piny air	breathe smell piny air
Every word and phrase he speaks is true	word phrase speaks true
The mule trod the treadmill day and night	mule trod treadmill night
Will you please answer that phone?	will please answer phone
Dots of light betrayed the black cat	dots light betrayed cat
The good book informs of what we ought to know	book informs ought know
The facts don't always show who is right	facts always show right
They took their kids from the public school	took kids public school
Non-native-accented/noise-vocoded sentences: Chapters 4 and 5	
Full sentence	**Keywords**
It snowed, rained, and hailed the same morning.	snowed hailed same morning
Note closely the size of the gas tank.	note closely size tank
Mend the coat before you go out.	mend coat before out
What joy there is in living.	what joy there living
Lift the square stone over the fence.	lift square stone over
The rope will bind the seven books at once.	rope bind seven books
Hop over the fence and plunge in.	hop over fence plunge
The friendly gang left the drug store.	friendly gang left store
Mesh wire keeps chicks inside.	wire keeps chicks inside
The frosty air passed through the coat.	frosty air through coat
A saw is a tool used for making boards.	tool used making boards
The wagon moved on well-oiled wheels.	wagon moved oiled wheels
Cars and busses stalled in snow drifts.	cars buses stalled snow
The set of china hit the floor with a crash.	set china floor crash
A yacht slid around the point into the bay.	yacht slid point bay
The horn of the car woke the sleeping cop.	horn woke sleeping cop
The heart beat strongly and with firm strokes.	heart strongly firm strokes
The Navy attacked the big task force.	navy attacked big force
The hat brim was wide and too droopy.	hat brim wide droopy
The grass curled around the fence post.	grass curled around post
He lay prone and hardly moved a limb.	lay hardly moved limb
The slush lay deep along the street.	slush lay deep along
A wisp of cloud hung in the blue air.	wisp cloud hung air
A pound of sugar costs more than eggs.	pound sugar more eggs
The fin was sharp and cut the clear water.	fin sharp cut clear
The term ended in late June that year.	term ended June year

A tusk is used to make costly gifts.	tusk used make gifts
Ten pins were set in order.	ten pins set order
Oak is strong and also gives shade.	oak strong gives shade
Add the sum to the product of these three.	sum product these three
The hog crawled under the high fence.	hog crawled under high
Move the vat over the hot fire.	move vat hot fire
Leaves turn brown and yellow in the fall.	leaves turn brown fall
The flag waved when the wind blew.	flag waved wind blew
Burn peat after the logs give out.	peat after logs out
Hemp is a weed found in parts of the tropics.	hemp weed found tropics
A lame back kept his score low.	lame kept score low
The boss ran the show with a watchful eye.	boss ran show watchful
The slang word for raw whiskey is booze.	slang raw whiskey booze
It caught its hind paw in a rusty trap.	caught paw rusty trap
The wharf could be seen at the farther shore.	wharf seen farther shore
Feel the heat of the weak dying flame.	feel heat weak flame
The tiny girl took off her hat.	tiny girl took hat
Pluck the bright rose without leaves.	pluck rose without leaves
Two plus seven is less than ten.	two plus seven less
The glow deepened in the eyes of the sweet girl.	glow deepened eyes girl
Clothes and lodging are free to new men.	clothes lodging free men
We frown when events take a bad turn.	frown events take bad
Guess the results from the first scores.	guess results first scores
A salt pickle tastes fine with ham.	salt tastes fine ham
The spot on the blotter was made by green ink.	spot blotter made ink
Mud was spattered on the front of his white shirt.	mud spattered front shirt
The cigar burned a hole in the desk top.	cigar burned hole desk
The empty flask stood on the tin tray.	empty flask stood tray
A speedy man can beat this track mark.	speedy man beat mark
He broke a new shoelace that day.	broke new shoelace day
The coffee stand is too high for the couch.	coffee stand high couch
She sewed the torn coat quite neatly.	sewed torn coat neatly
The sofa cushion is red and of light weight.	sofa cushion red light
At that high level the air is pure.	high level air pure
Drop the two when you add the figures.	drop two add figures
There was a sound of dry leaves outside.	there sound leaves outside
Torn scraps littered the stone floor.	torn scraps littered floor
Add the store's account to the last pence.	add account last pence
Add the column and put the sum here.	column put sum here
There the flood mark is ten inches.	there flood mark inches
The tongs lay beside the ice pail.	tongs lay beside pail

They could laugh although they were sad.	they laugh although sad
Farmers came in to thresh the oat crop.	farmers came thresh crop
The lure is used to catch trout and flounder.	lure catch trout flounder
Float the soap on top of the bath water.	float soap top water
A blue crane is a tall wading bird.	crane tall wading bird
A fresh start will work such wonders.	start work such wonders
He wrote his last novel there at the inn.	wrote last novel inn
Even the worst will beat his low score.	even worst beat score
The cement had dried when he moved it.	cement dried when moved
The fly made its way along the wall.	fly way along wall
Live wires should be kept covered.	wires should kept covered
The large house had hot water taps.	large house hot taps
It is hard to erase blue or red ink.	hard erase blue ink
Write at once or you may forget it.	write once may forget
The doorknob was made of bright clean brass.	doorknob made clean brass
The wreck occurred by the bank on Main Street.	wreck occurred bank street
A pencil with black lead writes best.	pencil lead writes best
The blind man counted his old coins	blind counted old coins
They took the axe and the saw to the forest.	took axe saw forest
The ancient coin was quite dull and worn.	ancient coin dull worn
Jazz and swing fans like fast music.	jazz fans like music
Rake the rubbish up and then burn it.	rake rubbish up then
Slash the gold cloth into fine ribbons.	slash cloth fine ribbons