The benefit received from visual information when listening to clear and degraded speech in background noise

Catherine Louise Blackburn

A thesis submitted in partial fulfilment of the requirements of Nottingham Trent University for the degree of Doctor of Philosophy

Nottingham Trent University

May 2019

Abstract

In order to improve speech understanding in background noise, visual information is used to enhance the incoming auditory signal. This enhancement is known as the visual speech benefit. Variation in the amount of visual speech benefit that is received by participants is the focus of this research project and is examined for both clear and vocoded speech. Vocoded speech simulates the type of speech experienced by cochlear-implant users. Experiments 1 and 2 examined variation in the amount of visual speech benefit gained if the type of background noise in the speech test changed. The key results from Experiment 1 and 2 were that the visual information provided was not enough to enhance speech understanding for particularly unintelligible speech. Experiment 3 assessed change to levels of visual speech benefit if the target talker in the stimuli changed. Significant differences in intelligibility between talkers was found. The amount of visual speech benefit increased as the audio intelligibility of the target talker decreased in clear speech. Overall, therefore, it is important that consideration is given to the levels of intelligibility provided by the stimuli used in speech perception testing as this may change outcomes. Experiments 4 and 5 examined individual differences that may predict the amount of visual speech benefit gained. In Experiment 4, the significant predictors of the amount of visual speech benefit gained in clear speech were general speech perception ability, ability to detect audio and visual synchrony, and tendency towards autistic traits. The results of Experiment 5 showed that general speech perception ability and time spent looking at the mouth area measured using eye-tracking were significant predictors of the amount of visual speech benefit gained in clear speech. Individual differences between participants may therefore predict differences in speech perception and should also be considered when testing speech perception.

Contents

Figures

Tables

Acknowledgments

I would like to thank my supervisor Paula Stacey for her help and support and my second and third supervisors, Pádraig Kitterick and Gary Jones.

Thank you to all my friends and colleagues at NTU for their participation and help with this project. Special thanks to Roy for his technical support, and Russell and Chris for their support.

Finally, I would like to thank my family. My children, Grace, Finlay and Tilly for their love and support, and my husband Nigel for his love and support and for believing in me.

Publications

Blackburn, C. L., Kitterick, P. T., Jones, G., Sumner, C. J., & Stacey, P. C. (2019). Visual Speech Benefit in Clear and Degraded Speech Depends on the Auditory Intelligibility of the Talker and the Number of Background Talkers. *Trends in Hearing, 23,* 1-14

Conference Presentations

- Blackburn, C.L. (2018, May). *Individual differences in the amount of benefit obtained from visual speech information when listening in noise.* Oral presentation at Psychology PhD and DPsych Conference, Nottingham Trent University.
- Blackburn, C. L., Kitterick, P. T., Jones, G., & Stacey, P. C. (2018, April) *Looking and Listening: Individual Differences in the amount of benefit obtained from visual speech information when listening in noise*. Poster presented at the Doctoral Research Festival, Nottingham Trent University.
- Blackburn, C.L. (2018, February). *Individual differences in the amount of benefit obtained from visual speech information when listening in noise*. Oral presentation at School of Social Sciences Research Conference, Nottingham Trent University.
- Blackburn, C. L., Kitterick, P. T., Jones, G., & Stacey, P. C. (2017, September). *Understanding Speech: Variability in the benefits obtained from visual speech information when the audio intelligibility of the talker varies.* Poster presented at BPS East Midlands Branch Conference, University of Derby.
- Blackburn, C. L., Kitterick, P. T., Jones, G., & Stacey, P. C. (2017, September). *Individual differences in the amount of benefit obtained from visual speech information when listening in noise*. Poster presented at Basic Auditory Science Annual Conference, University of Nottingham.
- Blackburn, C. L., Kitterick, P. T., Jones, G., & Stacey, P. C. (2017, June). *Looking and Listening: Variability in the benefits obtained from visual speech when listening to clear and degraded speech in background noise.*

Poster presented at British Society of Audiology Annual Conference, Harrogate.

- Blackburn, C. L, (2017, May). *Looking and Listening: Variability in the benefits obtained from visual speech when listening to clear and degraded speech in background noise*. Oral presentation at Psychology PhD and DPsych conference, Nottingham Trent University.
- Blackburn, C. L., Kitterick, P. T., Jones, G., & Stacey, P. C. (2016, September). *The enhancement of speech perception obtained from including visual information when listening to clear and degraded speech in different types of background noise.* Poster presented at Basic Auditory Science Annual Conference, University of Cambridge.

Prizes

Best Poster Prize (2017, September). Awarded at British Psychology Society East Midlands Annual Conference, University of Derby.

Chapter 1 – Overview of the Thesis

This chapter gives a brief introduction to the project and an overview of each of the thesis chapters.

1.1 Introduction

Understanding speech in quiet is a relatively easy task to achieve for normal hearing listeners, however, the presence of background noise provides a significant challenge to speech perception. In order to improve understanding, visual information is used to enhance the incoming auditory signal. This enhancement is known as the visual speech benefit. Variation in the amount of visual speech benefit that is received by participants is the focus of this research project. Significant variation in the amount of benefit received has been found in previous research but the reasons for this variation are largely unexplained. This project examines the possible reasons for variation in the amount of visual speech benefit gained by making changes to the speech that is to be understood and by assessing differences between individuals participating in the research.

Variation in the amount of visual speech benefit received is examined if the speech stimuli changes, by varying the type of background noise, or by varying the target talker. The first three experiments examine changes to the stimuli used in the research. Individual differences between participants that change the amount of visual speech benefit received are also examined. Experiments 4 and 5 examine individual differences between participants that are taking part in the research.

Variation in the amount of visual speech benefit gained is examined for both clear speech and for vocoded speech. Vocoded speech simulates the type of speech experienced by cochlear-implant users and has a different quality to clear speech.

Below is a summary of each chapter in the thesis.

1.2 Overview of the Thesis Chapters

Chapter 2 – Cochlear Implants

This chapter provides an overview of cochlear implants, which restore a sense of sound to individuals who are profoundly deaf. To simulate the experience of cochlear-implant users, normal speech is processed by vocoding and is used in research. This is also discussed in Chapter 2.

Chapter 3 - Audio, Visual only and Audio-Visual Speech Perception

Chapter 3 outlines the concept of speech. It then goes on to discuss auditory only speech perception and the challenges to this, including the particular difficulties understanding speech in the presence of background noise. Differences in speech perception for hearing-impaired populations are discussed, with particular reference to cochlear-implant users. The chapter also examines how visual and auditory information are combined to improve speech understanding. The benefit that visual information contributes to speech perception is then discussed. The details of this research project are outlined with reference to the current research assessing variation in visual speech benefit. Finally, the importance of this project is discussed.

Chapter 4 - Experiments 1 and 2

Chapter 4 considers the first two experiments in this project, which assess differences in the amount of visual speech benefit gained if the background noise in the experiment changes. Research suggests the amount of visual speech gained by participants may vary if the type of background noise changes. Experiment 1 uses a single target talker to assess levels of visual speech benefit received in background noise types of one, two and sixteen talkers. This is assessed for both clear and vocoded speech. The results of this experiment inform Experiment 2, which repeats the experimental paradigm of Experiment 1 but with changes to the target stimuli. Location information is added to the target stimuli by the inclusion of an interaural delay. This enables participants to more easily identify the target talker from the background noise.

Chapter 5 - Experiment 3

This chapter reports Experiment 3, changing the target talker. Research assessing the audio intelligibility of talkers has found variation between talkers, it is possible this variation may change the amount of visual speech benefit participants are able to gain from the talker. Initially, eleven different target talkers are assessed to establish differences in the levels of audio intelligibility using IEEE sentences (Rothauser, 1969) in a sixteen-talker background noise type. Four of these talkers, two male and two female, with the highest and lowest levels of audio intelligibility are taken forward to the main experiment. Levels of visual speech benefit are assessed for clear and vocoded speech for these four differing target talkers in a sixteen-talker background noise type.

Chapter 6 - Experiment 4

Experiment 4 examines whether the levels of visual speech benefit received by participants are predicted by individual differences between the participants. Previous research has assessed a variety of individual differences between participants that have changed the ability to understand speech and benefit from visual information. This experiment assesses whether some of these individual differences change the level of visual speech benefit gained. This is examined for both clear and vocoded speech. One single male target talker is used, informed by the results of Experiment 3. Individual difference measures are selected informed by previous research which employed audio only, visual only or audio-visual stimuli. Measures tested are; general speech perception ability, sensitivity to temporal fine structure information (TFS), aspects of attention from the Test of Everyday Attention (TEA) (Robertson et al, 1994), ability to detect audio-visual synchrony (AVS), verbal working memory capacity (Wechsler, 2008), and scores on the Autistic Spectrum Quotient (AQ) (Baron-Cohen, Wheelwright, Skinner, Martin & Clubley, 2001). The relationship between results on these tests and levels of visual speech benefit received are examined.

Chapter 7 - Experiment 5

Chapter 7 examines the final experiment in this project, Experiment 5. The significant predictors of visual speech benefit from the previous experiment are re-tested to establish if these are also predicted by gaze behaviour, as where participants are looking during the speech task may be affecting the amount of visual speech benefit gained. Performance in vocoded speech tasks was more varied than clear speech tasks in the previous experiments. A familiarisation task is included in Experiment 5 in order that performance levels in vocoded speech are more stable.

Therefore, Experiment 5 repeats the speech perception tests and measured the significant predictors of the amount of visual speech benefit gained from Experiment 4, with the addition of a familiarisation task to stabilise performance in the vocoded speech conditions and with the inclusion of eye tracking to assess gaze direction and duration during speech perception tasks. This is assessed for both clear and vocoded speech.

Chapter 8 – Comparison of Results across Experiments

This chapter compares some of the results across different experiments to further assess the variation in amount of visual speech benefit received. Comparison of the target talker used in Experiments 1 and 2 is made with the different target talkers used in Experiment 3. Comparisons are made to establish if the target talker used in Experiments 1 and 2 is more or less intelligible than the other talkers. Different results may have been found in Experiments 1 and 2 if a different target talker had been used in the stimuli. Secondly, the results of Experiments 4 and 5 are combined for clear speech and the resulting data discussed. Although the results of Experiment 5 were not significant, unlike the results of Experiment 4, it is possible combining the data from the two experiments may reveal significant patterns in the data.

Chapter 9 - Summary and General Discussion

The final chapter in this project, Chapter 9, provides a general summary of the thesis and the results of the experiments. The implications of the research are examined and suggestions for future research based on the results are discussed.

Chapter 2 – Cochlear Implants

This chapter provides an overview of how cochlear implants work and discusses vocoder simulation of a cochlear implant.

2.1. Cochlear Implants (CI)

A cochlear implant is a small electronic device that can restore a sense of sound to someone who is profoundly deaf. The external microphone receives the auditory signal which is turned into a digital code. The speech processor transmits this code to the receiver under the skin. The receiver converts the code to electric impulses and sends them along the electrode array, which has been surgically implanted into the cochlea. This stimulates the nerve and a signal is transmitted through the auditory system (Figure 2.1).

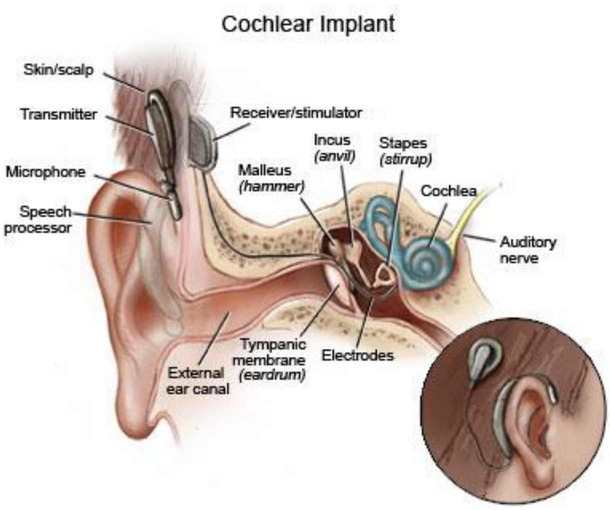


Figure 2.1 Diagram of a cochlear implant (O'Reilly & Morlet, 2016).

The signal is divided into frequency bands using bandpass filters in the speech processor. The bandpass filters extract the speech envelopes and transmits this as a series of rapid pulses. The pulses are delivered to a different electrode in the array, situated at a different point along the cochlea. Figure 2.2 shows a diagrammatic representation of a four-channel cochlear implant and the associated auditory signal in wave form. In this diagrammatic example, the incoming auditory signal is filtered into four bands and four pulses are generated onto four electrodes inserted into the cochlea. The rapid pulses delivered along the electrode array in an implant mirrors the tonotopic organisation of the basilar membrane in typical hearing individuals;

high frequency pulses are transmitted to the base and lower frequency signals to the apex. The pulses are increased in amplitude and length according to the intensity of the incoming auditory signal. It is important to note that the signal received by the auditory system is qualitatively different to that experienced by those with typical hearing. The signal has little TFS information and lacks spectral detail and therefore contains less information.

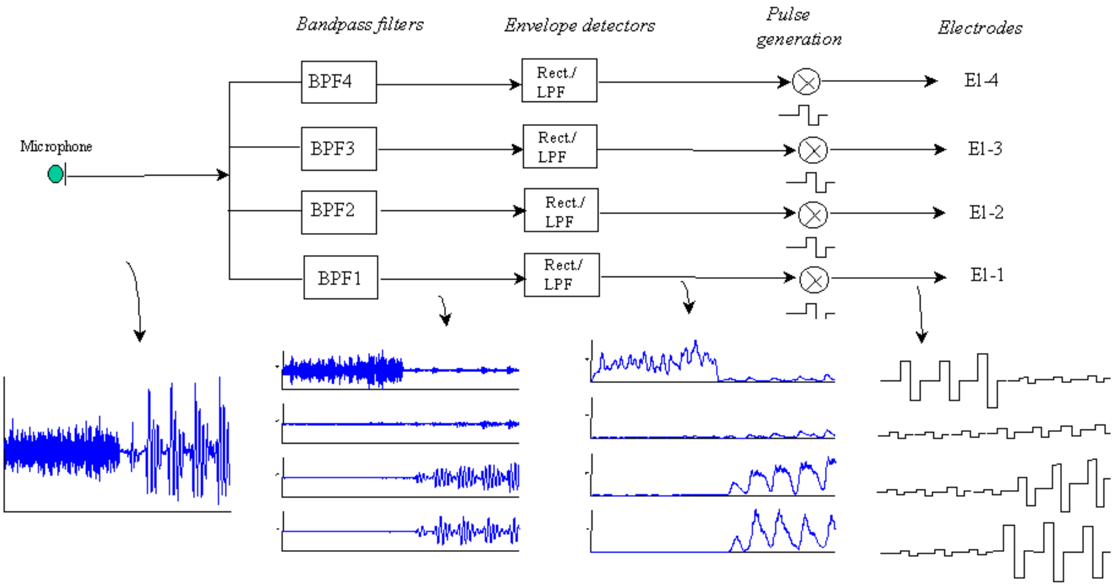


Figure 2.2 A schematic of a four-channel cochlear implant (Loizou, 1997) © 1997 IEEE

The number of electrodes inserted typically range from 12 to 24 but can be more or less than this. Figure 2.3 shows an x-ray image of a 29 electrode implant inserted into the human cochlea. The number of electrodes varies by manufacturer, as does the design of the implant, which is aimed at producing the most effective restoration of hearing. A larger number of channels does not typically equate to better hearing outcomes. Zeng et al. (2015) compared a newly designed 26 electrode implant but found no significant difference in performance levels between the 26 electrode implant and performance levels achieved by other research studies assessing performance levels of CI's with between 12 and 24 electrodes.

Insertion of the implant is a major surgical procedure with many factors to consider to ensure the most successful outcome for patients (Dhanasingh & Jolly, 2017). For example, the electrode array insertion depth and the make and model of implant to be inserted; the implants have different electrode lengths and degrees of flexibility.

The choice of implant is decided by the needs of the patient and the anatomy of the cochlea. Patients typically have different anatomical requirements in addition to the location and amount of cochlea damage. It is also important to minimise further trauma to the cochlea where possible to ensure the best audiological outcomes for patients (O'Connell, Hunter & Wanna, 2016).

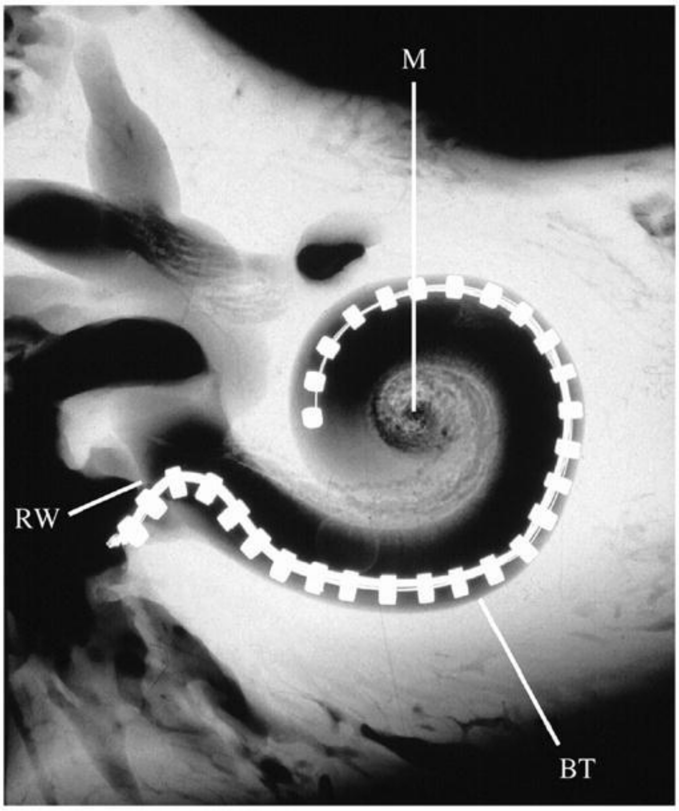


Figure 2.3 An X-ray image of a 29 electrode cochlear implant in a human cochlea. From Géléoc & Holt, 2014. Reprinted with permission from AAAS.

2.1.1 Variations in Outcomes for Cochlear-implant Users

While some individuals are able to successfully attain a reasonable level of hearing from their implants, others have a less successful outcome (Moberly, Bates, Harris & Pisoni, 2016; Pisoni, Kronenberger, Chandramouli & Conway, 2016). Rumeau et al. (2015) found that following implantation 35% of CI users were still unable to use the telephone, with a further 29% only able to use the phone with a familiar speaker and a familiar topic. Only 35% of the sample were able to fully use the telephone without constraints. Similarly, Lenarz, Sönmez, Joseph, Büchner and Lenarz (2012) conducted a review of the long-term outcomes for 1005 CI users. They found 13%

were classified as poor performers, scoring less than 10% on speech recognition tests. In contrast, 44% achieved 90-100% correct scores on the speech tests. This highlights the significant variation in speech perception following implantation.

As highlighted, there is significant variation in performance following implantation with some individuals achieving near perfect speech recognition in quiet conditions (Lenarz et al., 2012), however, in background noise outcomes are poorer. Turner, Gantz, Vidal, Behrens and Henry (2004) compared performance levels between normal hearing (NH) listeners and cochlear-implant users in a sixteen talker background noise. They found that the CI users had poorer performance levels than NH listeners in background noise, with a difference between them of more than 30dB. Nelson, Jin, Carney and Nelson (2003) found that CI users could identify approximately 80% of key words correctly in quiet conditions. However, in a steady-state background noise performance was significantly poorer. At a signal to noise ratio (SNR) of 8dB, comparable to everyday environmental background noise levels, performance dropped by approximately 50%. This evidences the significant detriment to performance caused by background noise to CI users.

The increased challenge presented by background noise is argued to be caused by the reduced information the CI provides. Much of the spectral and temporal information within speech is not transmitted by the implant, opportunities for ‘glimpsing’ and listening in the dips between speech are lost, therefore speech within background noise becomes more difficult to understand (e.g. Friesen, Shannon, Baskent & Wang, 2001; Fu, Shannon & Wang, 1998). This is discussed in more detail in section 3.3.1.

Reasons for differences in performance levels between CI users vary. Better outcomes have been shown for individuals who have residual hearing and who have used a hearing aid prior to implantation (Lazard et al., 2012). Additionally, a longer duration of deafness before fitting of the CI leads to poorer outcomes (e.g. Blamey et al., 1996; Blamey et al., 2013; Holden et al., 2013; Leung et al., 2005). Green at al. (2007) found the duration of deafness to be an independent predictor of CI outcomes, accounting for 9% of the variability in performance levels. Holden et al. (2013) also found age at implantation to be a significant predictor of performance, with increased age leading to significantly poorer outcomes. Holden et al. (2013) argue this is a result of cognitive decline with aging, leading to diminished speech perception ability. Cognitive abilities have been argued to lead to variation in speech perception abilities in all age groups. Differences between individuals in cognitive abilities are discussed in more detail in section 6.2 where individual differences in speech perception are examined.

Extensive research has been carried out to understand the differences in outcomes experienced by cochlear-implant users and so improve outcomes for patients. However, the substantial variation between CI users means it is problematic to compare performance between CI users. Therefore, research has been conducted on normal hearing listeners using speech that has been processed to simulate the type of speech heard by cochlear-implant users. This is discussed further in the next section.

2.2 Vocoder Simulation

Simulation of the type of speech sounds produced by a cochlear implant has been produced in order that research can be carried out (e.g. Shannon, Zeng, Kamath, Wygonski & Ekelid, 1995). This allows for studies using CI simulated speech to be carried out on normal hearing (NH) participants, increasing the participant pool and allowing for comparison across CI users and NH listeners. Testing with normal hearing listeners allows control over some of the potential confounding variables that may be present in CI users as discussed in section 2.2.4. Any observations from the vocoded speech can then be attributed to the novelty of the speech type (Fu, Nogaki & Galvin, 2005) or from the changes to the speech because of the vocoding (Fu & Shannon, 1999), and not due to functional changes within the perceptual systems of CI users.

How Speech is Vocoded

Vocoded speech is created by dividing the speech signal into a limited number of frequency bands and the speech envelopes extracted. The bands are recombined to produce the speech signal. This maintains the envelope cues in the speech but largely removes the TFS information. Vocoded speech can be noise vocoded (Shannon et al., 1995) or sine-wave vocoded (Dorman, Loizou & Rainey, 1997). Noise vocoded speech is made by using the envelope from each channel to modulate a noise band of the same width as the channel (Shannon et al., 1995). Sine-wave vocoded speech is created in a similar way, except the channels are used to modulate a sine wave at the centre of the channel (Dorman et al., 1997).

Vocoded speech reduces the spectral and temporal information available in the signal, which mirrors the experience of CI users. The number of channels in the vocoded speech signal can be varied to change the spectral detail of the signal, and the temporal resolution of the speech can be varied by changing the temporal cut-off frequency. In a cochlear implant, spectral and temporal information is typically processed independently; the number of stimulated electrodes is varied to change the spectral detail, and the cut-off frequency of the envelope filters and rate of stimulation of the electrodes is varied to change the temporal frequency. Using vocoder simulations, spectral and temporal information can be manipulated to assess the contribution these make to speech understanding for normal hearing participants and imitate the experience of CI users.

Increasing the number of channels and increasing the spectral content, is argued to improve speech intelligibility. In their seminal work, Shannon et al. (1995) examined speech perception for 1, 2, 3 and 4 bands of information for recognition of consonants, vowels and simple sentences, using noise vocoded speech. Speech recognition performance increased for all stimuli types as the number of bands of information increased, with near perfect performance levels with four bands of information. Research has suggested that speech recognition performance does not improve if more than eight channels of information are available for vowel identification (Dorman et al., 1997). Moreover, for sentence identification, an easier task than vowel identification, no significant improvement in performance was found if more than five channels of information were available (Dorman et el., 1997). However, cochlear implants typically have more than 8 channels of information as more are needed to ensure sufficient stimulation of the auditory nerve fibres. This is balanced with the possibility of adjacent channels causing interference. If they are spaced too closely, sound distortions can be caused.

More channels of information are required to understand speech if it is within background noise (Hopkins, Moore & Stone, 2008; Qin & Oxenham, 2003; Shannon et al., 1995). Qin and Oxenham (2003) tested speech perception using H.I.N.T sentences, which are simple high context sentences, with background noise types of modulated speech-shaped noise, steady-state speech shaped noise, a single male talker and a single female talker. They examined the speech reception threshold (SRT) required to achieve a 50% correct performance level using various processing conditions; unprocessed speech and speech with 4, 8 and 24 channels of information. Compared to clear speech, all of the processed speech background noise types resulted in an increase to the SRT required to understand the speech. Processed speech with 24 channels was easier to understand than 8 channels. The condition with 4 channels of information produced the highest SRT's and therefore the worst performance. Performance was also worse for the vocoded speech in the background noise with completing talkers, male and female, than the modulated or steady-state noise. This is argued to be as a result of the speech vocoding removing the TFS information from the speech signal which is particularly important for speech perception in noise, especially when there are competing talkers in the background. This is examined in more detail in Chapter 3 where the effects of background noise on speech perception are discussed.

Sine wave vocoded speech is argued to be more intelligible than noise vocoded speech (Chang & Fu, 2006; Fu, Chinchilla & Galvin, 2004; Gonzalez & Oliver, 2005; Souza & Rosen, 2009; Whitmal III, Poissant, Freyman & Helfer, 2007). Gonzalez and Oliver (2005) compared the intelligibility of noise vocoded speech and sine wave vocoded speech. They asked normal hearing participants to identify the target talker from a choice of two speech simulations, and to identify the gender of a target talker. The sine-wave speech showed a significant advantage in all conditions, it was easier to identify the gender of the target talker in sine wave vocoded speech and easier to identify the target sentence in sine wave vocoded speech. Souza and Rosen (2009) found sine-wave vocoded speech was more intelligible than noise vocoded speech but only with a 300 Hz envelope cutoff frequency. With a 30 HZ cutoff frequency, the opposite result was found, noise vocoded speech was more intelligible than sine wave vocoded speech. This is argued to be a result of better cues to periodicity (repetition of the signal waveform) and its variation in fundamental frequency (pitch) in sine-wave vocoded speech with high envelope cut-off frequencies (Souza & Rosen, 2009).

Sine-wave vocoded speech is argued to be a better representation of the hearing experience of cochlear implant users (Dorman et al., 1997). Whitmal III et al. (2007)

argue sine wave vocoded speech produces a better approximation of speech envelope fluctuations present in speech and therefore better characterisers the experience of CI users. Sine wave vocoded speech can therefore be used by researchers investigating the effects of CI speech on normal hearing listeners (e.g. Stacey, Kitterick, Morris & Sumner, 2016).

Limitations of Vocoder Studies

Despite the benefits of using vocoded stimuli in research, it has been argued there are some differences to the experience of CI users. It has been argued that CI users experience a distorted sound owing to limitations of electrode insertion in the basilar membrane during surgery (Fu & Shannon, 1999; Ketten et al., 1998). In addition, hearing loss is argued to lead to changes to the neurological structure of the brain, meaning audio signals may be processed differently in hearing impaired than in normal hearing participants. For example, Lazard and Giraud (2017) argue that reorganisation of the visual cortex following deafness leads to compromised phonological processing following the fitting of a cochlear implant.

Adaptation to vocoded speech has also been shown over time by CI users. Most improvement in performance is argued to take place over the first three months of use (e.g. Spivak & Waltzman, 1990), however, adaptation has also been shown over longer time periods (Tyler, Gantz, Woodworth, Fryauf-Bertschy & Kelsay, 1997). This suggests CI users may exhibit better performance than normal hearing listeners because they are experienced with this type of sound. Therefore, comparison across hearing impaired and normal hearing participants should make allowances for the level of experience gained by participants with vocoded stimuli.

Despite these concerns, it is important that research can be carried out to assess the likely performance levels that can be achieved by CI users. Using vocoded stimuli as a simulation for the type of speech experienced by cochlear-implant users allows control over possible confounding variables, particularly the significant variation in performance levels. Therefore, using vocoded stimuli remains an important resource within research. This research project uses sine-wave vocoded speech to simulate the speech experience of CI users.

2.3 Chapter Summary

- Profound hearing loss can be in part remediated by the fitting of a cochlear implant
- Variation in outcomes from cochlear implants is significant. Research examining causes of this variation is therefore important
- Vocoded simulations mimic the type of sound experienced by cochlear-implant users, and despite some concerns, is an important tool for research into cochlear implant hearing using normal hearing populations.

Chapter 3 – Audio, Visual and Audio-Visual Speech Perception

Chapter 3 first outlines the concept of speech. It then examines speech perception for normal hearing populations. This is discussed with reference to auditory only speech perception, and the particular difficulties faced when understanding speech in the presence of background noise, and visual only speech understanding. The chapter then examines differences in speech perception for hearing-impaired populations; the challenges of background noise and the use of visual information. Audio and visual integration to benefit speech understanding is then considered. Finally, the benefit that visual information provides to auditory speech perception is assessed, and current research examined. Details of this project are then outlined, with reference to the importance of this research area.

3.1 Characteristics of Speech

Speech is defined as the oral communication that humans use to express thoughts and feeling to each other. It involves the complex movement of muscles in the head, neck, chest and abdomen to produce specific decodable sounds. Speech develops gradually over the first few years of life and is learned through observation and repetition.

The smallest unit of sound in speech is a phoneme. Phonemes are combined to produce complete words. For example, the words hat is divided into three phonemes, ‘h’ ‘a’ and ‘t’. Phonemes are different to letters as they represent the sounds of language and not the written text. There are 44 phonemes in the English language that are combined to produce over one million different words. Around 20,000 words are used commonly by an English speaker. Despite this extensive vocabulary, speech production and understanding are relativity easy tasks for humans to accomplish.

Speech is created by vocal fold vibration filtered through the vocal tract which produces complex waveforms. Speech is therefore a complex sound, divided into multiple frequency bands. In contrast to a simple sine wave, which has a regular repeating pattern, speech is a made up of multiple sine waves. Figure 3.1 shows a complex wave form pattern for the IEEE sentence (Rothauser et al., 1969), ‘We must vote in the election next Tuesday’. The complex speech signal is characterised by rapid fluctuation in the temporal fine structure information (TFS), as shown in Figure 3.1. TFS information is particularly important for speech understanding, and is largely removed by a cochlear implant or vocoder simulation (as discussed in sections 2.2.3 and 2.2.5), making speech understanding more challenging.

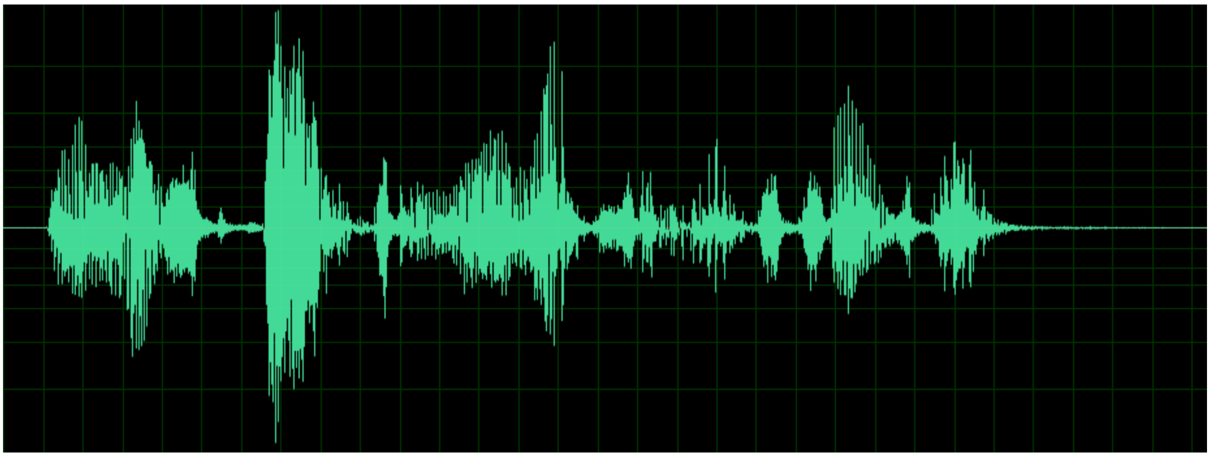


Figure 3.1 A complex speech wave for an IEEE sentence.

3.2 Audio and Visual Speech Perception in Normal Hearing Populations

3.2.1 Auditory Speech Perception

Despite the complexity of speech sounds, the understanding of speech remains a relatively simple task for normal hearing listeners. During speech production the characteristics of the talker vary significantly, and therefore the speech signal varies significantly. The physical features of the talker, such as tongue position (Marin, Pouplier & Harrington, 2010) and vocal tract length variability (Fitch & Giedd, 1999) are known to change the auditory output. The speech can also be modified by other characteristics of the talker such as age, gender, accent, speaking rate and emotional state (Weatherholtz & Jaeger, 2016). Speech has also been shown to be more difficult to understand if the talker is unknown to the listener, in contrast to someone the listener is familiar with (Johnsrude et al. 2013; Souza, Gehani, Wright & McCloy, 2013). Despite these complexities, speech perception remains a relatively effortless task. This is supported by specialist brain regions dedicated to speech understanding.

It has been argued that speech has special representational properties in the brain that differentiates it from other auditory input (Hickok, & Poeppel, 2007). In addition to the recognised areas of specialisation, Broca’s area, the angular gyrus, the insular cortex and Wernicke’s area, it has been argued that speech comprehension is supported by additional specialism in the anterior temporal pathway (Scott, Blank, Rosen, & Wise, 2000). These areas work in combination to perceive and express language.

The perception of speech does not rely solely on the incoming auditory information stream but also uses cognitive resources to interpret and understand the incoming speech signal. For example, in his seminal research, Warren (1970), replaced words in a sentence with a cough or a 1000Hz tone. Nineteen out of the 20 participants reported that all of the speech sounds were present in the cough condition and all 20 participants reported hearing all of the speech sounds in the tone condition. This type of phonemic restoration demonstrates the importance of cognitive processes when perceiving speech. Additionally, other cognitive processes such as memory and learning have been shown to contribute to speech perception (Davis & Johnsrude, 2007). Speech perception can also be supported by the use of context, semantic cues and syntactic cues to aid understanding of the content (Miller & Isard, 1963). This is particularly important in background noise, where the auditory target may be masked by other extraneous auditory information (Pollack, 1975).

3.2.2 Challenges to Auditory Speech Perception - Listening in Noise

Research demonstrates that speech perception becomes more challenging when the target audio stream is in competition with additional auditory information (Miller, 1947). The target speech is masked by the extraneous auditory information. Therefore, understanding speech in noise is a complex activity, requiring auditory processing and the use of cognitive resources to disentangle the target speech from the background auditory stream of information.

The neural processes involved in understanding speech in noise have been argued to be greater than listening to speech in quiet. This has been shown for active speech tasks (Golumbic et al., 2013) and for passive listening tasks (Evans, McGettigan, Agnew, Rosen & Scott, 2016). The ability to understand speech in noise has also been argued to decline with age but can be moderated by cognitive abilities, such as memory and attention, and central processing speed of speech in the brainstem (measured by subcortical responses to speech syllables) (Anderson, White-Schwoch, Parbery-Clark & Kraus, 2013). In addition, socioeconomic status, linked to higher educational attainment, has been shown to account for a significant proportion of the variation in speech in noise performance levels (Anderson et al., 2013). This suggests the ability to perceive speech in noise is modifiable through life experience, and varies with the characteristics of the listener, therefore, individual variation in performance levels could be expected. This is discussed in Chapter 6 and 7 in this project where individual differences between participants are discussed in detail.

Speech understanding in background noise is argued to be more challenging as the target speech is ‘masked’ by the background noise. Masking takes the form of energetic masking and informational masking. Energetic masking occurs when the target and background noise overlap in time and frequency. This makes it more difficult to hear the target speech within the external noise (Pollack, 1975). Informational masking refers to any other type of masking, and typically occurs when the signal and the masker are both audible but share similar characteristics, therefore, listeners are unable to disentangle the two incoming streams of information perceptually (Brungart, 2001; Pollack, 1975; Shinn-Cunningham, 2008). For example, energetic masking can occur if white noise masks the target speech if a high signal to noise ratio (SNR) is used; the white noise is louder than the target speech and so masks the speech signal. In contrast, if the background noise consists of another talker or talkers, some energetic masking may take place, but informational masking will typically have greater masking effects. The semantic and syntactic qualities of the target speech and background speech are shared and therefore the speech and background are perceptually similar, and the content is masked.

Variations in Masking Effects

Background noise can be steady state, whereby the spectral and temporal qualities are generally constant, such as a machine ‘hum’. This type of noise typically has less limited spectral content and therefore does not mask speech unless the spectral details of the target speech and background noise overlap. In contrast, fluctuating or modulated noise has variation in both the spectral and temporal information, such as a background noise of other people talking, and therefore can have greater masking effects.

Masking effects can vary as a result of ‘glimpsing’ of the target speech (Cooke, 2006). These ‘dips’ in the speech signal of the masker are argued to allow small windows in the spectro-temporal audio information stream that provide opportunities to infer part of the complex speech sound. Spectral dips arise where the frequency content of the target and background source differ, allowing brief glimpses of the target speech. Similarly, temporal dips arise where there are brief pauses in the speech or during low energy speech sounds such as m, n, k or p.

Masking effects can also vary as a result of the stimulus used in the study. For example, Schoof and Rosen (2015) found that high sentence predictability increases the benefits from glimpsing in the speech. Schoof and Rosen (2015) compared speech perception performance using an adaptive procedure for BKB sentences (high predictability; Bench, Kowal & Bamford, 1979) and IEEE sentences (low predictability; Rothauser et al., 1969). Increased benefit from dip listening was found for the BKB sentences, suggesting the amount of masking experienced by listeners may depend on the content of the speech information. Therefore, speech understanding in noise maybe improved if the speech content is less complex.

Reduction in masking effects has also been argued to take place if priming is given for the target talker. Freymann, Balakrishnan and Helfer (2004) found that hearing a preview of the target voice, prior to sentence trials, primed listeners to recognise and ‘latch on’ to the target voice more easily. This reduced the masking effects of the competing voices as attention was directed to the target voice more readily.

Increased masking effects are also argued to take place if the background talker is in the same language as the target talker. A reduction in informational masking is argued to take place if the competing streams are of different languages. Kilman, Zekveld, Hällgren and Rönnberg (2014) found that increased English language proficiency in native Swedish speakers lead to increased masking for Swedish speech in English background babble. Low proficiency in English led to a release from informational masking for the Swedish target sentences.

Overall, it can be seen that background noise presents a significant challenge to speech understanding. In order to improve understanding, additional perceptual resources are utilised. In many situations, listeners can see as well as hear their communication partners, the visual information provided is available as an aid to understanding. This is discussed in detail in the next section.

3.2.3 Visual Speech Perception

Visual information is defined as being able to see the talker's mouth, including the lips, tongue and teeth (Peelle & Sommers, 2015), and other facial movements such as the jaw. These facial movements serve to provide phonetic and temporal cues to aid the perception of the target speech. Mouth movements serve as an additional cue as to when we should attend to the talker (Picou, Ricketts & Hornsby, 2011). Additional visual cues from the talker have also been shown to aid understanding, such as gestures (Morrel-Samuels & Krauss, 1992) and facial expressions (Adolphs, 2002). Research has also suggested visually impaired people have difficulty learning certain phonemes and make more speech production errors across phoneme boundaries (Mills, 1987). This suggests visual information is important in the development of speech perception and production, evidencing the benefit obtained from visual speech information. Similarly, visual information has been shown to take on more importance when trying to understand complex text and for heavily accented speech, even when the speech is clear (Reisberg, McLean & Goldfield, 1987).

Typically, visual only speech perception (also known as lip reading or speech reading) is considered to be a difficult task, resulting in poor performance. For example, Middelweerd and Plomp (1987) found performance levels of 6.4% for older adults and 23.1% for younger adults lip reading high context meaningful sentences. Lip reading for single consonants resulted in slightly better performance levels. Grant, Walden and Seitz (1998) found performance levels for lip reading consonants ranged from 21% to 40% correct, with a mean performance level of 31%.

As well as being more challenging than auditory only tasks, visual only tasks result in greater variability in performance (e.g. Macleod & Sumerfield, 1987). For example, for normal hearing children lip reading sentences with high context information, performance levels varied between 0% and 41% correct (Lyxell &

Holmberg, 2000). Similarly, in a normal hearing adult population, testing everyday sentences, such as ‘here’s a nice quiet place to rest’, performance levels varied from 0% to 65% correct (Auer & Bernstein, 2007). Task performance has also been shown to vary as a result of the task stimuli. In contrast to Auer and Berstien (2007), Altieri, Pisoni and Townsend (2011) tested word recognition using open set sentences with less contextual information (CUNY sentences) and found mean word recognition scores per sentence to be 12.4%, much lower than when using high context sentence stimuli.

This large variation in lip reading skills is argued to be largely unexplained (Summerfield, 1987) and are not significantly improved with training (e.g. Dodd, Plant & Gregory, 1989). Tests over the life span have suggested that visual only performance declines with age (e.g. Sommers, Tye-Murray & Spehar, 2005). Tye-Murray, Spehar, Myerson, Hale and Sommers (2016) tested normal hearing adults with ages ranging from 22 to 92 years. Performance on closed set sentences varied from less than 10% correct to over 90% correct across the age groups. For the age group 22-30, mean performance was nearly 70% correct, this declined through each age group, reaching the lowest level of approximately 35% correct at age 81-92. Feld and Sommers (2009) argue poorer lip reading skills demonstrated by older adults, may partly be explained by declining short-term working memory capacity (Park et al., 2002) and processing speed (Eckert, 2011), which are known to reduce with age. There is some evidence that differences in the speed of low level neural processing generally, may account for variation in lip reading ability in all age groups (e.g. Shepherd, DeLavergne, Frueh, & Clobridge, 1977).

As highlighted, visual information is important for understanding speech, particularly within background noise, where understanding of speech is significantly disrupted, as discussed in section 3.1.3 (Miller, 1947). Therefore, the ability to lip read may contribute to the amount of visual speech benefit participants are able to receive; the visual and auditory information are combined to improve speech perception. This integration of auditory and visual information to benefit speech understanding is discussed in section 3.4.

3.3 Audio Only and Visual Only Speech Perception in Hearing Impaired Populations

3.3.1 Audio Speech Perception in Hearing Impaired Populations

In hearing impaired (HI) listeners auditory information is absent or compromised. Individuals with a hearing impairment have been shown to have particular difficulty understanding speech in background noise (e.g. Festen & Plomp, 1990; Peters, Moore & Baer, 1998). This is a particular issue for cochlear-implant users, and is argued to be as a result of a loss of temporal and spectral information (e.g. Baer, Moore & Gatehouse, 1993; Bernstein & Oxenham, 2006). In addition, temporal fine structure information (TFS) is argued to be an important contributor to understanding speech in noise (Eaves, Summerfield & Kitterick, 2011; Moon & Hong, 2014). TFS information is can be compromised in listeners with a hearing impairment and therefore speech understanding in noise is more challenging. The lack of access to TFS information also compromises the ability to benefit from the 'dips' in the fluctuating background noise (e.g. Feston & Plomp, 1990; Hopkins, Moore & Stone, 2008).

It has also been argued HI listeners have difficulty selecting auditory objects and filtering out background noise and attending to the target speech using selective attention and rapid attention switching (see Shinn-Cunningham, 2008). In addition, the cognitive load experienced by HI listeners when processing and understanding the speech signal, is argued to be greater than experienced by NH listeners, resulting in less resource available to exploit any dips in the speech signal to improve speech perception (Shinn-Cunningham, 2008).

It has also been argued difficulties perceiving speech in noise can be experienced by individuals, despite the absence of clinical hearing loss. This has been termed obscure auditory disfunction (OAD) (Saunders & Haggard, 1989), auditory disability with normal hearing (King & Stevens, 1992) or King-Kopetzky syndrome (Hinchcliffe, 1992). Badri, Siegel and Wright (2011) compared speech in noise performance of listeners with OAD to normal hearing listeners. They presented BKB sentences (Bench et al., 1979) in speech shaped noise at SNR's of -5, -8 and -10dB. Significantly poorer sentence recognition was found for the OAD group at the lowest SNR, -10dB. A mean score of 26% correct was found for OAD group compared to

41% correct for the NH group. Badri et al. (2011) found the OAD group had significantly wider auditory filters than the NH group. This is argued to be a result of peripheral cochlea damage, for example to the hair cells in the cochlea, or possibly a result of impairment in higher central functioning, such as the descending auditory pathways.

Cochlear-implant Users

As outlined in section 2.2.3, to restore a sense of hearing to profoundly deaf individuals, a cochlear implant can be fitted. The quality of sound experienced by cochlear implant (CI) users is different to normal hearing listeners, the sound is degraded spectrally and temporally. Additionally, much of the temporal fine structure information (TFS) is removed and therefore noisy environments present a particular challenge (Lorenzi, Gilbert, Carn, Garnier & Moore, 2006). This is argued to be a result of the reduced opportunities for release from masking effects experienced by CI users (Friesen, Shannon, Baskent & Wang, 2001; Fu, Shannon & Wang, 1998). Nelson, Jin, Carney and Nelson (2003) found that CI users performed significantly worse than normal hearing participants in a speech perception task in background noise. This was shown for steady state and for modulated background noise, with no difference in performance levels for the CI users at 16dB. At an SNR of 8dB, a slight release from masking effects in the modulated noise was found. Additionally, when compared to sentence recognition in quiet, performance levels for the CI group also showed a significant reduction in performance. At an SNR of 16dB, performance levels for the CI group dropped by more than 20%, and at an SNR of 8dB dropped by more than 50%. Nelson et al. (2003) argue this demonstrates that CI users are unable to benefit from release from masking in modulated noise, even at favourable SNR's. Additionally, the same lack of masking release did not vary as a result of the type of CI used, suggesting that the effect is a general characteristic of CI processing, demonstrating the lack of spectral information provided by the CI.

Similarly, Stickney, Zeng, Litovsky and Assmann (2004) tested CI users speech in noise performance using a background noise of a single competing talker and, secondly, steady state noise. In comparison to the normal hearing group, CI user's performance levels declined more sharply with the addition of background noise.

They also found poorer performance levels with a single competing background talker when compared to steady state noise, suggesting a detriment to performance from both informational and energetic masking effects was taking place. This was in contrast to the normal hearing group, who demonstrated a release from masking effects in the competing talker condition. The normal hearing group were able to benefit from the temporal dips in the fluctuating background noise, unlike the CI listeners, who were not able to benefits from the dips in the background noise.

Vocoded Speech

Audio only speech perception has also been examined using vocoded speech and normal hearing listeners. As outlined in section 2.2., speech can be processed to simulate the type of speech experienced by cochlear-implant users; vocoded speech. Using vocoded speech stimuli has allowed comparison of speech perception performance between CI users and normal hearing (NH) listeners using the same type of stimuli. Differences or similarities between these two groups can then be attributed to the speech type or the possible effects of the hearing loss. For example, Stickney et al. (2004) observed similar lack of release from masking using a CI group and a vocoded speech group, suggesting masking effects were caused by the speech type and not as a result of any differences between the groups as a result of physiological hearing loss.

Nelson et al. (2003) used vocoded stimuli to compare performance levels across NH and CI users. They found the vocoder simulation group were affected by a lack of release from masking in background noise in a similar way to the CI group. They used four channel vocoded speech and found no benefit was gained from the dips in the modulated noise at 8dB and very little gain at 16dB. The temporal gaps in the background speech were as long as 500ms. This suggests vocoded speech provides a useful facsimile of CI speech and is masked by background noise in a similar way.

As discussed above, the particular difficulty understanding speech in background noise experienced by CI users has been widely investigated. This difficulty leads to the benefit gained from visual information to take on even greater importance. This is examined in the next section.

3.3.2 Visual Information for Hearing Impaired Populations

As considered in the above section, the challenges to speech perception for hearing impaired (HI) populations are typically greater than normal hearing (NH) listeners. Similarly to normal hearing populations, visual information is used by HI populations as an aid to understanding (Erber, 1975; Kaiser, Kirk, Lachs, & Pisoni, 2003). However, as a result of the compromised or absent auditory information, visual information can take on more importance for speech perception in HI individuals than for NH individuals. As a result, some hearing impaired populations have been shown to process visual information differently to normal hearing populations.

Research has compared lip reading abilities between normal hearing and hearing impaired populations. Hearing impaired populations have been shown to have enhanced speech reading abilities (Auer & Bernstein, 2007; Bernstein, Tucker & Demorest, 2000). Bernstein et al. (2000) compared performance levels between normal hearing participants to those of hearing impaired participants for nonsense syllables, single words and sentences. In all conditions, the HI sample outperformed the NH sample. Similarly, Auer and Bernstein (2007) found HI adults identified 44% of target words correctly, compared to 19% correct for the NH group. The HI samples in Bernstein et al. (2000) and Auer and Bernstein (2007) had pre-lingual hearing loss and therefore gained speech reading expertise through experience with visual information without any auditory input. This contrasts with other research (e.g. Ronnberg et al. 2013, Summerfield, 1992) who found no speech reading advantage for HI populations and whose population sample were largely post-lingually deafened adults. This suggests early hearing loss contributes to enhanced speech reading skills, which are not attained to the same degree with later onset hearing loss (Pimperton, Ralph-Lewis and MacSweeney, 2017).

This is further supported by more recent research by Tye-Murray, Hale, Spehar, Myerson and Sommers (2014). Results of speech reading tests showed that speech reading ability improved between 7 and 14 years of age, with age accounting for more than 35% of the variance in speech reading ability between NH and HI groups. Tye-Murray et al. (2014) argue this may reflect increased practice with visual information due to impoverished hearing. Alternatively, they suggest enhanced speech reading ability may develop in conjunction with phonological reading skills, that are typically developed at this age. Tye-Murray et al. (2014) go on to argue that it may also reflect enhanced speech reading skills that are acquired pre-lingually and then maintained throughout adulthood.

Pimperton, Ralph-Lewis and MacSweeney (2017) also found enhanced speech reading abilities for cochlear implant (CI) users over NH adults. They assessed speech reading ability using single words as the test stimuli. The results showed significantly better performance levels were achieved by the CI users; the mean number of words correctly identified was 22 for the NH group and 41 for the CI group. There was also a significant positive correlation between age at implantation and speech reading ability; a higher score on the test correlated with later implantation. This suggests that a longer duration of hearing loss and so greater dependence on visual information, led to enhanced ability to use the visual information. This again suggests early auditory deprivation leads to enhanced speech reading skills. Auditory input, in the form of a cochlear implant, reduces the reliance on visual information and speech reading skills are not improved further.

Overall, it can be seen that for individuals with hearing loss, particularly those fitted with a cochlear implant, access to visual information is especially important as an aid to speech perception. The nature of the speech signal experienced by CI users and the neurological changes that are argued to take place once hearing loss is experienced change the way speech perception is experienced compared to normal hearing individuals. Access to visual information is especially important in background noise, as masking effects present a challenge to speech understanding. The integration of the auditory and visual information is argued to differ in HI populations to NH populations. Multisensory integration, with particular reference to speech perception, is discussed in the next section.

3.4 Multisensory Integration for Normal Hearing Populations

3.4.1 Defining Multisensory Integration

Multisensory integration refers to the process by which information from different sensory systems is combined in the brain and a weighted average of inputs produced to facilitate a coherent understanding of the world. Integration enables alteration in perception through combination, enhancement, depression or compensation of the senses. Input from multiple senses is a constant and necessary function in everyday life yet remains largely unconscious. Although an evolutionary adaptive process, it is also argued integration is an ability that develops through experience. Stein, Stanford and Rowland (2014) argue neurons in new born infant's brain are not capable of multisensory integration. This ability develops and grows through cross modal experience in the environment. Integration is argued to be largely beneficial, for example, it allows increase in speed of response (Diederich & Colonius, 2004) and enhanced stimulus detection (Lovelace, Stein & Wallace, 2003). However, the sensory input from multiple sources can also be misleading. For example, the ventriloquist illusion demonstrates how visual and auditory information are combined into a single percept, although transmitted from different locations (Alais & Burr, 2004). This illusion also demonstrates how the auditory and visual systems are readily integrated in normal hearing listeners.

Audio-visual integration for normal hearing listeners, involves the fusing of heard and seen information into unified percept. It is argued to be an innate process, evidenced by a predisposition to integrate sight and sound. Wertheimer (1961) found infants moved their head towards an audible click shortly after birth, and similarly, Kezuka, Amano and Reddy (2017) found the ability to localise sound using vision develops in accuracy from age 4 months to 7 months. The sound-induced flash illusion also demonstrates the innate fusion of auditory and visual information. In this illusion, the number of auditory beeps heard by participants inflates the number of flashes to be counted during the task (Shams, Kamitani & Shimojo, 2002). Using this illusion, Nava and Pavani (2013) found children were more susceptible to the effects of the auditory information at aged 6-7, the auditory beeps increases the number of visual flashes perceived. The effect reduced through the ages until visual information dominance was reached at age 11-12 and the effects of the illusion reached similar levels to those found in adults. This research evidences how multisensory integration may change in typically developing children.

The primary research area in assessing the integration of audio and visual information relates to speech perception. This is discussed in detail in the next section.

3.4.2 Models of Audio-Visual Speech Perception

It is argued that auditory and visual information are combined early in processing to enable multisensory integration of speech information. These 'amodal' or 'modaility neutral' accounts propose speech perception involves auditory and visual modalities from the beginning of the speech perception process (e.g. Summerfield, 1992). In contrast, other researchers argue the audio and visual information streams are initially analysed separately and integrated at a later stage (Grant, Walden & Seitz, 1998) (see Rosenblum (2008) for a review). Finally, a third model proposes that both early and late integration takes place, which may explain why both the early and late stage models are not fully supported by the current research evidence (Peelle & Sommers, 2015). Figure 3.3 shows the three models of audio-visual speech integration; early, late and multistage integration.

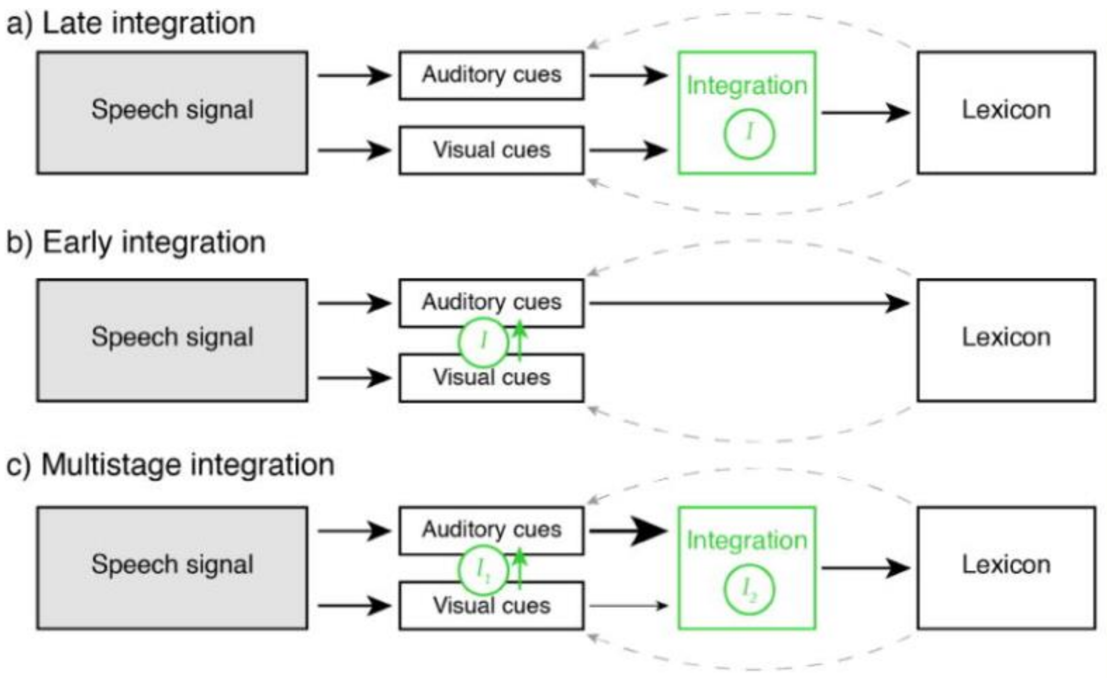


Figure 3.3 Models of audio-visual speech perception; a) late integration model, audio and visual information is processed separately and integrated later b) early integration model, audio and visual information are processed simultaneously, c) multistage model, allows for both early and late integration of audio and visual information (Peelle & Sommers, 2015).

An early proponent of the early integration model, Summerfield (1992) argued integration takes place before phonetic categorisation of the speech and therefore is integrated very early in the perceptual process. Summerfield (1992) suggests this is shown by the benefit of lip reading in background noise; visual information supplements the audio information before understanding of the speech is achieved. Secondly, Summerfield (1992) argues when the speech is replaced by a tone, so carries no phonetic information, speech can be more easily understood than with the visual information alone. Rosen, Fourcin and Moore (1981) replaced the auditory speech information with a voice pitch tone, which coordinated with the opening and closing of the vocal folds of the talker. Improvements to speech understanding were found when the tones were played. This suggests the auditory information in the form of a tone, was adding to the visual speech understanding before the speech content was understood. Summerfield (1992) also argued speaking rate changed the perception of speech tokens. Green and Miller (1985) found that the perception of bi/pi changed as a function of the speed that the visual information was provided. This suggests that the audio perception of the speech changed as the speech rate changed and therefore the visual information changed the speech perception. The audio and visual information were combined before categorisation of the speech took place.

The early influence of visual information on auditory perception is also evidenced by the McGurk effect (McGurk & MacDonald, 1976). In the McGurk effect, a visual stimulus and auditory stimulus show incongruent information, which leads participants to hear a new sound. For example, a talker saying ‘ga’, and a simultaneous audio stimulus saying ‘ba’, may lead participants to hear ‘da’. This demonstrates how the perceptual system integrates the audio and visual information into a unified percept before speech understanding is reached. The McGurk effect has been used in research extensively to demonstrate the early integration of audio and visual information (e.g. Green, Kuhl, Meltzoff & Stevens, 1991; Munhall, Gribble, Sacco & Ward, 1996).

Supporters of the late model argue integration takes place after processing of the audio and visual information independently (Grant, Walden & Seitz, 1998) (Figure 3.3). For example, in the McGurk effect (McGurk & Macdonald, 1976) large asynchronies between audio and visual stimuli still result in the effect taking place.

Similarly, Massaro, Cohen and Smeele (1996) found that when simple syllable stimuli were presented to participants with offset times between audio and visual information of up to 500ms, there was little disruption to the integration of the audio and visual information. Massaro et al. (1996) argue this supports the Fuzzy Logic Model of Perception (Massaro, 1989), whereby audio and visual information is evaluated independently prior to integration. Figure 3.3 shows a schematic of this model. Speech information is first analysed as separate audio and visual information streams, integrated, and finally recognition or decision about the content of the speech is made (Massaro, 1989).

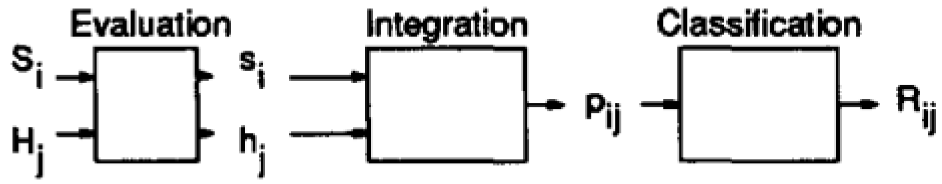


Figure 3.4 Diagram of the Fuzzy Logic Model of speech perception (Massaro, 1989).

Finally, the multistage model of integration argues that both early and late integration takes place (Peelle & Sommers, 2015) (Figure 3.3). Peelle and Sommers (2015) argue that visual information alters the processing of auditory information as it is being processed, at an early perceptual stage, as evidenced by crossmodal cortical activity during speech perception. Additional speech gestures, (such as articulation) are argued to affect later integration and improve speech perception. Both early and late integration is argued to be dependent on the information available to the listener and is therefore a dynamic and flexible process.

Overall, it can be seen that in normal hearing populations, the integration of audio and visual information enhances speech understanding. However, hearing impaired populations are argued to perceive speech in a different way and therefore integrate information differently. This is discussed in detail in section 3.4.4.

3.4.3 Audio-Visual Integration of Speech and Neurological Areas

Words that are ambiguous in the auditory or visual modality, can be correctly identified in a combined audio-visual modality; degraded signals can be supported by information from the other modality and overlap of the lexical neighbourhood of the words in the visual and auditory modality achieved to support speech understanding (Tye-Murray, Spehar, Myerson, Hale & Sommers, 2016).

Many of the neurological areas employed to understand audio only speech are also employed in perceiving visual information when understanding speech (Irwin et al., 2017). Visual information is argued to be relayed early in the integration process, through multiple pathways and on to the auditory system to aid speech understanding (Peelle & Sommers, 2015). It has also been argued that visual information has a special status for enhancing audio speech; it is processed in the auditory system rather the visual system as would be expected of other visual input (Bernstein & Liebenthal, 2014). This is demonstrated in the fMRI scans shown in Figure 3.5 (adapted from Calvert, et al. 1997). The pink voxels show areas of activation when participants were shown visual only speech and the blue voxels show activation for audio only speech. The yellow voxels show the overlapping areas activated by both audio and visual speech in the lateral temporal auditory cortex. The audio areas activated by the visual speech were not activated when non-speech facial movement was shown to the participants. This demonstrates the unique effect of visual speech information on auditory pathways in the cortex.

Additionally, where audio input is degraded, the visual system is argued to show enhanced neurological activation (Sekiyama, Kanno, Miura & Sugita, 2003). This is discussed in detail in sections 4.1 and 4.2.2, where the contribution of the visual information is argued to be enhanced by degraded audio input; the Principle of Inverse Effectiveness.

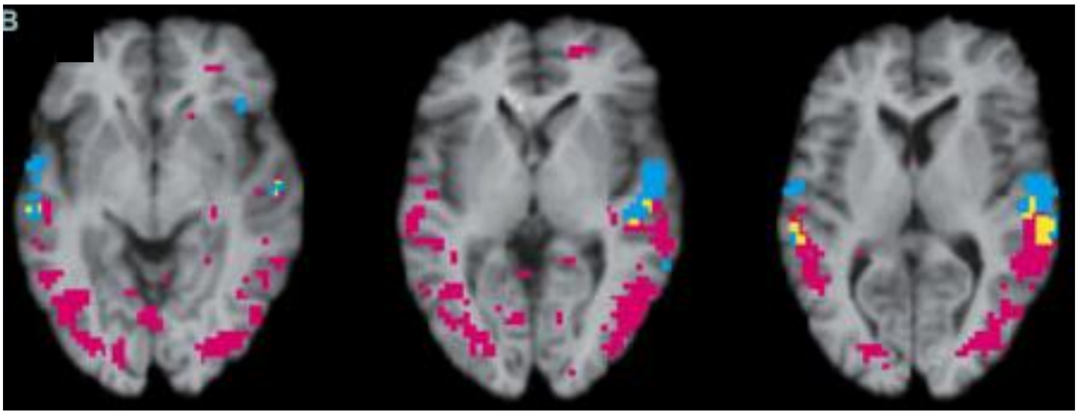


Figure 3.5 fMRI scans of brain activation during speech perception. The pink voxels show areas of activation when participants were shown visual only speech, the blue voxels show activation for audio only speech, the yellow voxels show the overlapping areas activated by both audio and visual speech (adapted from Calvert, et al. 1997). Reprinted with permission from AAAS.

3.4.4 Multisensory Integration for Hearing Impaired Populations

It has also been argued that the neural changes brought about by hearing loss change the way multisensory integration is achieved by individuals with a hearing impairment. ERP evidence suggests there is more efficient use of visual speech cues by individuals with hearing impairments (Winneke & Phillips, 2011). Winneke and Phillips (2011) argue visual information is used more effectively to compensate for the suboptimal auditory processing experienced with hearing loss acquired through normal aging processes.

Cortical reorganisation has also been shown to take place across auditory and visual systems for cochlear-implant users. Rouger et al. (2012) found that during hearing loss, brain regions used for face and voice processing are reallocated to assist in visual only speech tasks. Following cochlear implantation, these brain areas are reorganised to allow more efficient audio-visual integration of speech (Anderson, Wiggins, Kitterick & Hartley, 2017; Chen, Puschmann, & Debener, 2017; Rouger et al., 2012). Areas engaged in visual only processing are reactivated to process faces and voices (Rouger et al., 2012). Similarly, Stropahl and Debener (2017) found evidence of cross-modal cortical reorganisation in participants with mild to moderate hearing loss. This suggests cortical reorganisation takes place very early on during hearing loss to enable individuals to enhance the visual information stream to support the auditory impairment. This is in contrast to normal hearing listeners who do not experience the same level of plasticity in audio and visual brain regions if they retain normal hearing.

It is therefore argued that hearing impaired populations may integrate auditory and visual information in a different way to normal hearing populations. Moreover, the benefit received from being able to see, in addition to being able to hear, the target talker may differ between normal hearing and hearing impaired listeners. This is examined in detail in the next section.

3.4.5 Visual Speech Benefit for Normal Hearing Populations

Speech perception is known to be improved by the addition of visual information to the auditory information stream (Erber, 1969; Grant & Seitz, 2000; MacLeod & Summerfield, 1987; Sumby & Pollack, 1954). Visual information has been shown to provide between four decibels (dB) (Middelweerd & Plomp, 1987) and 15dB (MacLeod & Summerfield, 1987) of benefit to the audio only information stream when presented in background noise for normal hearing populations. Middelweerd and Plomp (1987) tested younger adults, age between 19 and 28, and older adults, aged between 68 and 84 years, using sentences in speech shaped background noise. They reported a visual speech benefit of 4.6dB for the younger age group and a benefit of 4dB for the older age group. Similarly, MacLeod and Summerfield (1987) found an average visual speech benefit of 11dB for sentences in white noise, with the amount of benefit ranging from 6 to 15dB between participants. The significant variation in the amount of visual speech benefit gained between individuals has been examined in further research.

It is argued the amount of visual speech benefit may vary as a result of the type of stimuli used in the study. For example, if the background noise and target speech is presented at different signal to noise ratios (SNR's) (e.g. Ross, Saint-Amour, Leavitt, Javitt & Foxe, 2006). Variation in the SNR is argued to change the way in which the audio and visual information are integrated as a result of the effects of the Principle of Inverse Effectiveness (PofIE), whereby impoverished unisensory input leads to increased multisensory gain. This is discussed in more detail in sections 4.1 and 4.2.2.

Variability has also been observed for different types of speech stimuli, for consonant and vowel recognition and for words in sentences and in different types of background noise. Grant, Walden & Seitz (1998) found significant variation between participants with noise induced sensory hearing loss, testing consonant recognition and recognition of key words in low context IEEE sentences (Rothauser et al., 1969), in speech-shaped background noise. The amount of visual speech benefit varied significantly between individuals, with the average benefit reported as 44% (SD=17.8%), ranging from 8.5% to 83%. Sommers et al. (2005) found variation across participants testing vowels-consonant-vowels phrases, single words and meaningful sentences in a twenty talker background noise. Variation in performance levels was found between participants for consonants (proportion correct ranging from approximately 0.8 to 0.5), for single words (proportion correct ranging from approximately 0.8 to 0.5) and sentences (proportion correct ranging from approximately 0.75 to 0.2). Similarly, Van Engen, Xie and Chandrasekaran (2017) found significant variation in the amount of visual speech benefit gained by participants testing simple high context sentences in a two talker background noise. At –2dB, visual enhancement, calculated by taking the difference between the AV and AO performance and normalising it by the improvement available given the AO performance level, ranged from 1 (maximum score available, all key words correct in the AV condition) to -1. At -20, performance levels ranged from approximately 0.7 to -.05.

Variation in visual speech benefit has also been found across populations. Tye-Murray, Spehar, Myerson, Hale and Sommers (2016) examined variation in the amount of visual speech benefit gained across age groups. When unimodal performance in the audio only and visual only task were controlled for in the analysis, no difference across age groups was found in the amount of benefit received. Variation was therefore attributed to individual differences between participants.

Differences in visual speech benefit have also been observed for individuals with a diagnosis of autism. Stevenson et al. (2018) argue that multisensory integration is suboptimal in those with a diagnosis of an autism spectrum disorder (ASD). They found a wider temporal binding window (TBW) for ASD individuals which was specific to complex speech stimuli (McGurk stimuli), in contrast to more simple stimuli; flashes and beeps, dynamic handheld tools. Lower levels of visual speech benefit have also been found in individuals with an ASD (Smith & Bennetto, 2007). This is examined in detail in section 6.2.6.

Variation in gaze behaviour between individuals is also argued to predict the amount of visual speech benefit gained. Individuals that spent longer looking at mouth regions have been shown to receive significantly more visual speech benefit (Alsius, Wayne, Paré & Munhall, 2016; Rennig, Wegner-Clemens & Beauchamp, 2018). This is examined in detail in section 7.2.1.1.

Variation in the amount of visual speech benefit received is examined in detail in this project. This is examined if the stimuli changes; if the type of background noise varies and if the target talker changes. This is also examined with reference to individual differences between participants. Detailed discussion of this is provided in the chapters that follow. Variation in visual speech benefit is also examined using vocoded speech, simulating the experience of CI users. Variability in the amount of visual speech benefit gained for hearing impaired populations is discussed in the next section.

3.4.6 Visual Speech Benefit for Hearing Impaired Populations

Differences in the amount of visual speech benefit have also been explored for hearing impaired populations. It has been argued, greater levels of visual speech benefit are achieved by individuals with hearing impairments (see Stevenson, Sheffield, Butera, Gifford & Wallace, 2017 for a review). For example, Goh, Pisoni, Kirk and Remez (2001) compared the level of visual speech benefit of one CI user to 25 normal hearing participants and found increased visual speech benefit for the CI user over the normal hearing participants. In a larger sample of 20 CI users, Kaiser et al. (2003) found enhanced visual speech benefits in a single target talker condition for identification of single words, but not in a condition where the target talker changed between each trial. Kaiser et al. (2003) argue this suggests CI users have enhanced ability to extract individual talker information and apply this knowledge to subsequent trials to improve speech perception; more successful utilisation of visual information leads to greater levels of visual speech benefit.

Rouger et al. (2007) explored differences between CI users and normal hearing participants using vocoded speech. The longitudinal data of 97 CI users was compared to performance levels of normal hearing participants, who were tested using vocoded speech. Greater visual speech gain was found for the CI users over the normal hearing participants. Rouger et al. (2007) argue CI users develop specific audio-visual abilities to overcome auditory speech challenges; speech reading abilities are developed during deafness which contribute to improved audio-visual performance after implantation and enhanced multisensory integration. However, Stacey, Kitterick, Morris and Sumner (2016) argue that the superior visual speech benefit gain shown by the participants in Rouger et al. (2007) is attributable to the underperformance of the normal hearing population sample. Stacey et al. (2016) presented IEEE sentences (Rothauser et al., 1969) to participants in a multi-talker babble background noise. Greater levels of visual speech benefit were found in vocoded speech conditions than in clear speech conditions for all stimuli types; open set and closed set sentences. Performance levels for the normal hearing participants tested by Stacey et al. (2016) was compared to the performance of the normal hearing participants tested by Rouger et al. (2007). Modelling of this data suggests that the normal hearing participants in Rouger et al. (2007) were suboptimal in their integration of the audio and visual information compared to the normal hearing participants in Stacey et al. (2016), who displayed optimal integration. Stacey et al. (2016) go on to suggest the amount of visual speech benefit may vary as a result of changes to the type of background noise. The level of visual speech benefit gained may also vary depending on the type of target speech stimuli used. This may account for differences in the amount of visual speech benefit gained between studies comparing NH and CI users.

In summary, it can be seen that the amount of visual speech benefit may vary between normal hearing listeners and CI users. This project examines differences between clear speech and vocoded speech, which simulates the experience of CI users. Variation in the amount of visual speech benefit received can then be attributed to the type of speech used. Further details of the project are summarised in the following section.

3.5 Summary of this Research Project

Chapter 3 outlined the difficulties of perceiving speech within background and the particular challenges faced by CI users. The use of visual information in order to assist with speech perception has also been discussed. There is currently a lack of research examining variation in the benefit received from visual speech information when listening to speech in noise, notably for vocoded speech, which simulates the type of speech experienced by CI users.

This research project aims to enable better understanding of challenges faced by CI users by using vocoded speech within the speech stimuli. Variation in the amount of visual speech benefit gained by individuals is also examined if the type of background noise changes or if the target talker changes. Additionally, the large variation in the amount of visual speech benefit gained is examined by examining individual differences between participants. There are suggested individual differences that contribute to visual only, audio only and audio-visual speech task performance but it is not known whether these predict the amount of visual speech benefit received. Finally, individual differences in eye movements are examined to assess if eye gaze direction and duration change the levels of visual speech benefit gained by participants. There are known gaze patterns utilised when perceiving speech, which vary when background noise is present.

The results of this project will make an important contribution to the body of knowledge addressing the amount of benefit visual information provides to speech perception. In addition, the work using vocoded speech may allow better prediction of outcomes for individuals following CI fitting. This could enable increased accuracy in framing outcome expectations as well as improved support after implantation. This is important for CI users to maintain and enhance social connectedness and quality of life. Ultimately, the results may also be relevant to other hearing-impaired populations such as hearing aid users and could therefore be of wider benefit.

3.6 Chapter Summary

- Speech perception is a relatively easy task for humans to accomplish
- Background noise presents a particular challenge to speech understanding, due to the effects of the noise masking the speech signal
- In order to assist with speech understanding, visual information is used to enhance the speech signal
- Speech perception for hearing impaired populations, particularly cochlear-implant users, is especially demanding in background noise and therefore the visual information has greater importance
- The benefit received from the visual speech information to speech perception varies between studies and between individuals
- This project aims to address the variation in the amount of visual speech benefit received by examining changes to the test stimuli and individual differences between participants. This will be addressed for clear speech, and for vocoded speech (which simulates the type of speech experienced by cochlear-implant users).

Chapter 4 – Visual Speech Benefit in Different Types of Background Noise

Chapter 4 examines the first two experiments in this project; variation in the amount of visual speech benefit received if the number of talkers in the background changes. It is not known whether the amount of visual speech benefit received varies as a result of changes to the background noise type. This is assessed for clear and vocoded speech, with particular reference to the Principle of Inverse Effectiveness. Experiment 1 examines the variation in the amount of visual speech benefit received if one, two or sixteen talkers are used as the background noise type. Experiment 2 repeats the methodology of Experiment 1 but includes changes to the target stimuli to make the speech more intelligible within the background noise. Results are discussed in relation to comparisons between the background noise types and differences between the amount of visual speech benefit received between clear and vocoded speech.[1]

4.1. Introduction

Chapter 4 examines the first two experiments in this project. As outlined in Chapter 3, in order to enhance speech understanding, visual information is integrated with the incoming audio information. The integration of visual and auditory information is known to enhance speech perception (e.g. Middelweerd & Plomp, 1987; Sumby & Pollack, 1954). The amount of benefit the visual information adds to the auditory information is known as the visual speech benefit. Experiments 1 and 2, discussed in this chapter, assess the changes to the amount of visual speech benefit that is obtained by participants if the background noise used in the stimuli changes. Background noise is known to reduce speech understanding because of masking effects that take place. Masking effects are discussed in detail in section 3.1.3. The additional benefit of the visual information to the auditory information stream is argued to be of greater importance in background noise, as a result of masking effects.

[1] The results from Experiment 2 and Experiment 3 in this project were published in part in the following publication; Blackburn, C. L., Kitterick, P. T., Jones, G., Sumner, C. J., & Stacey, P. C. (2019). Visual Speech Benefit in Clear and Degraded Speech Depends on the Auditory Intelligibility of the Talker and the Number of Background Talkers. *Trends in Hearing, 23*, 1-14

There are two types of masking that inhibit speech understanding in background noise; informational masking and energetic masking. Informational masking is argued to be largely internal to the listener, the lexical content of the target and background noise may be similar, and the listener may attend to the background talker rather than the target talker. Energetic masking occurs largely in the external environment, where the target talker is masked by the energy of the background noise i.e. the target cannot be distinguished from the background noise. Different masking effects are argued to take place depending on the type of stimuli and background noise (Rosen, Souza, Ekelund & Majeed, 2013).

It has also been argued that the Principle of Inverse Effectiveness (PofIE) may apply to the integration of audio-visual stimuli and the amount of visual speech benefit received. In their seminal work, Sumby and Pollack (1954) argued that as speech intelligibility is reduced due to the addition of background noise, the contribution of visual information uniformly increases (the Principle of Inverse Effectiveness; PofIE). Visual information is therefore more important in noisy environments than in quiet. Similarly, where the audio stream of information is most degraded, additional information provided by the visual information stream may be at its most beneficial (Meredith and Stein, 1986). Therefore, if speech is degraded by vocoding (to simulate the type of speech experienced by cochlear-implant users), the contribution of the visual information would be at its greatest. The PofIE has been explored in research addressing the contribution of visual information to speech perception. This is discussed further in section 4.2.2.

Research using an audio only paradigm has highlighted changes to performance levels if the background noise changes. It is not known whether changes to the background noise type would also change the amount of visual speech benefit that individuals receive. Therefore, Experiments 1 and 2 outlined in Chapter 4 will examine the variation in the amount of visual speech benefit received if the background noise changes. Background noise types of one talker, two talkers and sixteen talkers are examined. This will also be discussed with reference to the PofIE. There is currently a limited range of research in this area.

This is assessed for both clear speech and for vocoded speech. Vocoded speech simulates the type of speech experienced by cochlear implants (CI) users. This is discussed in detail in section 2.2.4. Sine-wave vocoded speech is used in Experiment 1 and 2, and throughout the project. Clear speech is easier to understand for normal hearing listeners as it is more familiar and is more intelligible. Therefore, in Experiment 1 and 2 and throughout the project, it would be expected than performance in clear speech would be better than for vocoded speech. Normal hearing listeners are used as participants throughout the project.

4.2 Experiment 1 – Changes to the Background Noise

4.2.1 Variation in the Type of Background Noise

Research has suggested that performance in an audio only speech perception tasks varies with changes to the background noise type. Masking effects vary as a result of changes to the background noise, and therefore, performance in tasks vary. Generally, performance on audio only tasks has been shown to decline with the addition of a single background talker (e.g. Brungart, 2001), and from one background talker to two background talkers, and with the addition of more talkers thereafter (e.g. Brungart, 2001; Carhart, Johnson & Goodman, 1975; Humes, Kidd & Fogerty, 2017; Simpson & Cooke, 2005). However, performance does not decline uniformly with the addition of more talkers, as might be expected, but has some variation, depending on the task and stimuli used.

Brungart (2001) used co-ordinate response measures as the stimuli in his research looking at audio only speech masking. He argued that with one background talker, masking effects are predominantly caused by informational masking and not energetic masking. This is evidenced by the finding that at 0dB, when the target talker and background talker would have equal energetic masking effects, the target words or background words are reported in nearly all of the responses. This suggests the information from both streams was heard, but the background sentence was reported and not the target sentence. The listener attended to the incorrect stream of information, caused by informational masking.

Freyman, Balakrishnan and Helfer (2004) used nonsense sentences in clear speech, with background noise types of varying number of talkers. They found that performance was significantly worse in the two talker background noise than the one talker background noise. Performance then improved with three background talkers, declined with the addition of more talkers until six talkers was reached, and then plateaued after six talkers. The significant loss of performance in the two talker condition is argued to be as a result of increased effects of informational masking (IM) at this point. Attention is drawn three ways, so selecting the correct stream of information to attend to is more problematic, and increased lexical interference is also argued to be at its most effective at this point. As the number of talkers increased to three, release from IM is argued to take place, which then declines as further talkers are added to the background noise and energetic masking effects take over. In contrast, Cullington and Zeng (2008) also found different effects in clear speech using BKB meaningful sentences (Bench, Kowal & Bamford, 1979). They found performance levels decreased significantly from one to two talkers and from two talkers to three talkers, but found no change in performance levels from three to four talkers. Freyman et al. (2004) argue differences in the stimuli and the task may account for differences in the level of IM taking place across studies. For example, Freyman et al. (2004) used nonsense sentences where less semantic interference would take place than if meaningful sentences were used, such as the BKB sentences used by Cullington and Zeng (2008).

Rosen et al. (2013) used IEEE sentences in an audio only paradigm to assess variation in performance in differing types of background noise. They used clear and vocoded background noise and clear speech targets. Vocoding of the target and background noise were implemented separately and then mixed together. In their second experiment, for clear speech in a background noise of one, two and sixteen talkers, performance was measured using an adaptive track procedure. Performance was not significantly different between the three types of background noise, with signal-to-noise ratios (SNR's) of approximately -1.5dB at the 50% correct performance level. However, in their first experiment, fixed SNR's of -6 and -2 were used. Once combined, the results of these experiments showed a significant reduction in performance between one and two background talkers at the lower SNR, -6dB, with performance in the one talker background noise plateauing at -2dB. This suggests that as SNR's are increased, background noise of more than one talker has greater masking effects.

Rosen et al. (2013) also assessed performance using 12-channel noise-vocoded speech as the background noise and clear speech target sentences. As expected, performance was significantly worse in the vocoded speech condition than in the clear speech condition. Performance levels in the one, two and sixteen talker background noise were not significantly different in the vocoded speech, with SNR's of approximately 6-7dB's in all conditions. Rosen et al. (2013) argue that the non-significant differences across vocoded background noise types reflects the lack of glimpsing opportunities in vocoded speech. There are few opportunities for glimpsing in this type of speech with very few background talkers and therefore the addition of further talkers does not change this, floor effects are already reached. In contrast, research by Cullington and Zeng (2008) used eight channel sine-wave vocoded speech and normal hearing listeners, as used in this project, and found that as the number of background talkers increased from one to two, there was a significant decrease in audio performance, suggesting increased masking effects were taking place. There was no further change in performance levels from two to three talkers or with the addition of further talkers. This again suggests differences across research which may be attributed to the stimuli used and the task requirements. However, in contrast to the methodology used by Rosen et al. (2013), the target and background noise were mixed and then vocoded together by Cullington and Zeng (2008), which may also have changed performance levels. In addition, Rosen et al. (2013) used IEEE sentences, which have lower predictability than the BKB sentences used by Cullington and Zeng (2008). Listeners may have been better able to predict and attend to the BKB target sentences because they are less cognitively demanding than IEEE sentences, resulting in differing performance levels.

Overall, it can be seen that masking effects are dependent on the task requirements and stimuli used. There is some variation among research, as outlined above, as to what type of background noise and at what SNR that takes place. However, it can be seen that informational masking, rather than energetic masking, is at its greatest when there are smaller numbers of background talkers. It is not known whether changes to the background noise that change audio performance levels, would also change the levels of visual speech benefit received.

4.2.2 Principle of Inverse Effectiveness (PofIE)

Sumby and Pollack (1954) argued that as speech intelligibility reduces due to the addition of background noise, the contribution of visual information uniformly increases (the Principle of Inverse Effectiveness; PofIE). Visual information is therefore more important in noisy environments than in quiet. Similarly, if speech is degraded by vocoding, the contribution of the visual information would be at its most beneficial. Research examining the Principle of Inverse Effectiveness (PofIE) with regard to speech perception has been conducted.

Using a six-talker background noise, Tye-Murray, Sommers, Spehar, Myerson and Hale (2010) examined speech perception with reference to the PofIE. They found that for both younger and older adults, there was no evidence of enhanced visual speech recognition in more degraded speech conditions as would be expected if the PofIE were to apply. They argue that integration of audio and visual speech information fundamentally differs to the integration of other bimodal stimuli. Tye-Murray et al. (2010) go on to argue that the PofIE may not be a generalisable ‘principle’ but is task dependent, influenced by the context and content of the stimuli. This is in contrast to the monotonic increase in benefit from the visual input to the auditory information stream reported by Sumby and Pollock (1954). Similarly, Ross, Saint-Amour, Leavitt, Javitt and Foxe (2006) argue there is a ‘special zone’ at a signal to noise ratio of -12dB where maximal multi-sensory integration occurs. At -12dB, they argue, audio-visual performance is more than three times higher than audio performance alone; the sensory system is tuned to be most efficient at intermediate levels. Therefore, the PofIE does not apply monotonically to speech perception; gains do not increase in direct monotonic relationship as might be expected.

The differences between performance in Ross et al.’s (2006) research and that of Sumby and Pollack (1954) is argued to be due to the methodology used. Sumby and Pollack’s (1954) words lists were relatively small and presented to participants before the study. Ross et al. (2006) argue this produced artificially high results in visual speech benefit. In contrast to Ross et al. (2006), Altieri and Wenger (2013) found the highest levels of audio-visual gain were at a signal to noise ratio of -18dB. Altieri and Wenger (2013) argue this is again related to the choice of words used in the task, they used a smaller word list than Ross et al. (2006) and therefore ’constrained task difficulty’. This resulted in a maximal performance level being achieved at a lower dB. Ross et al. (2006) and Altieri and Wenger (2013) contest the choice of words used in the test stimuli produced the differences in results, however, it is possible that other facets of the stimuli differed and should be examined further. The background noise used was different in each study; Sumby and Pollack (1954), used noise derived from a gas-tube source, Ross et al. (2006) used pink noise, while Altieri and Wenger (2013) used white noise. There are known differences between these background noise types (Halley, 1996) which may significantly affect the levels at which maximum performance is achieved and may account for changes in levels of the speech benefit received.

More recent research has examined the amount of visual speech benefit individuals receive in noise consisting of background talkers. The levels of visual speech benefit received in clear speech and vocoded speech were also compared. Stacey, Kitterick, Morris and Sumner (2016) examined the levels of visual speech benefit received by participants in a multi-talker babble background noise. They found a marginal increase in the amount of visual speech benefit received in vocoded speech over clear speech. This is in line with the PofIE; where the audio stream of information is most degraded, additional information provided by the visual information stream may be at its most beneficial; as unimodal performance declines multi-sensory integration is improved (Meredith and Stein, 1986).

It is worth considering however, that in Stacey et al. (2016) and in the methodology used in this project, the SNR for each participant is measured using an adaptive track procedure and therefore the SNR varies for each participant. The variation in the SNR across participants may mean interpretation of the PofIE becomes problematic, as the SNR varies adaptively for each participant. An alternate method of measuring this would be to present stimuli at a constant SNR, as presented in the research of Ross et al. (2006) and Altieri and Wenger (2013).

Overall, it could be argued that if the PofIE were to apply, the addition of further background talkers would result in increased levels of visual speech benefit; greater number of talkers would result in greater levels of benefit. Similarly, as vocoded speech is more degraded than clear speech, greater levels of visual speech benefit would be found in vocoded speech over clear speech.

4.2.3 Summary of Experiment 1

To assess the variation in the amount of visual speech benefit received by participants, clear and vocoded speech conditions were tested. Background noise types of one, two and sixteen talkers were used. Research has shown that the largest change in performance is between one and two talkers in audio only research (e.g. Cullington and Zeng, 2008; Rosen et al., 2013). Increased levels of visual speech benefit have been found in vocoded speech conditions than clear speech conditions in a multi-talker babble (Stacey et al., 2016), similar to a sixteen talker background noise type.

4.2.4 Hypotheses for Experiment 1

It was hypothesised that clear speech would be easier to understand than vocoded speech, and that audio-visual performance would be better than audio only performance.

For clear speech and for vocoded speech it was hypothesised that audio only and audio-visual speech perception performance would decline with an increase in the number of background talkers.

It was also hypothesised there would be more visual speech benefit in vocoded speech than in clear speech, and that there would be a greater benefit from visual speech information as the number of background talkers increased.

4.3 Method

4.3.1 Design

A within-participants design was used in Experiment 1. Audio only (AO) and Audio-visual (AV) Speech Reception Thresholds(SRT_{50}), measured in decibels (dB) were obtained for clear and vocoded speech for each of the background talker types; one talker, two talkers and sixteen talkers.

The first analysis was a 2 (modality: audio only vs audio-visual) x 2 (speech type: clear vs vocoded) x 3 (background noise type, one, two and sixteen talkers) ANOVA. The dependent variable was SRT's measured in dB.

A separate condition measured performance in a visual only task and was scored as percentage correct.

4.3.2 Participants

The number of participants recruited for Experiment 1 was based on the number used in other research projects testing speech perception in a similar way, and the availability of participants at Nottingham Trent University. For example, Stacey, Kitterick, Morris & Sumner (2016) calculated an interaction effect of 0.38, based on their first experiment testing speech perception in multi-talker babble background noise. Stacey et al. (2016) assumed a within-participants correlation between auditory only and audio-visual performance of 0.5, and a power of 0.80 and an α of 0.05, indicating 16 participants would be required. This number was taken as the minimum number required for each experiment in this project, which also allowed for the availability of participants within Nottingham Trent University.

Thirty-two participants were recruited from staff and students at Nottingham Trent University for Experiment 1. All reported normal hearing and normal or corrected to normal vision and spoke English as their first language. Students were rewarded with research credits. Consent was obtained from each participant as agreed by the Nottingham Trent University Research Ethics Committee. The mean age of the participants was 22.26, age range 19-37, with seven being male. The adaptive track for each condition was considered to be successful if performance levels were between 30% and 70% correct. Data for six participants had overall levels of performance of between 10 and 20% in the audio only vocoded condition, indicating that the SRT would be unreliable in this condition; participants would have less than 50% correct performance. The data for these participants was therefore removed from the analysis. A further three participants were removed because of missing portions of data, leaving 23 participants.

4.3.3 Materials

4.3.3.1 Equipment

Experiment 1 was conducted in a quiet experimental room within the psychology department at Nottingham Trent University. The audio stimuli were presented at 70-73dB. This was calibrated by presenting the stimuli over headphones attached to an artificial ear (G.R.A.S. 43AA) and measured using a microphone (ACO 7052E) connected to a sound level meter (SVAN 977). Audio was played over HD280pro headphones (Sennheiser, Wedemark, Germany) via a custom built digital-to-analogue converter. Visual stimuli were presented on a computer monitor with a screen measuring 41cm x 26cm. Stimulus presentation was controlled using E-Prime software (Version 2.0 Psychology Software Tools Inc., Sharpsburg, US) and using Matlab programming (Mathworks, Nantick, US).

4.3.3.2 Target Stimuli

The sentences were chosen from the IEEE corpus (Rothauser et al, 1969). The IEEE corpus consists of 720 short sentences grouped into phonetically balanced groups of ten. Three hundred sentences were recorded for use in this experiment by a single male target talker.

The videos were recorded on a Panasonic AVC HD camera. The audio files were recorded using a Studio Series SL150 microphone. The camera was positioned 80cm in front of the talker, positioned at head height and in front of a plain white background. This allowed a full head view of the target talker with no other features visible in each video. An example of this is shown in Figure 4.1. The audio and video track recorded were processed using Adobe Premiere Pro CC (v9.2). The videos were processed as mp4 files, with a frame size of 1920 x 1080 for the main experiment, and as .m2v files for the practice task. The audio files were processed as wav files, with a sample type of 4410Hz. Each sentence lasted approximately three seconds. Examples of the sentences are (key words are underlined), ‘<u>Cars</u> and <u>buses</u> were <u>stuck</u> in <u>snow</u> <u>drifts</u>’ and ‘<u>Use</u> a <u>pencil</u> to <u>write</u> the <u>first</u> <u>draft</u>’. Twenty-three of the sentences were used in the practice for the main experiment, and a further 260 sentences were used in the main experiment.

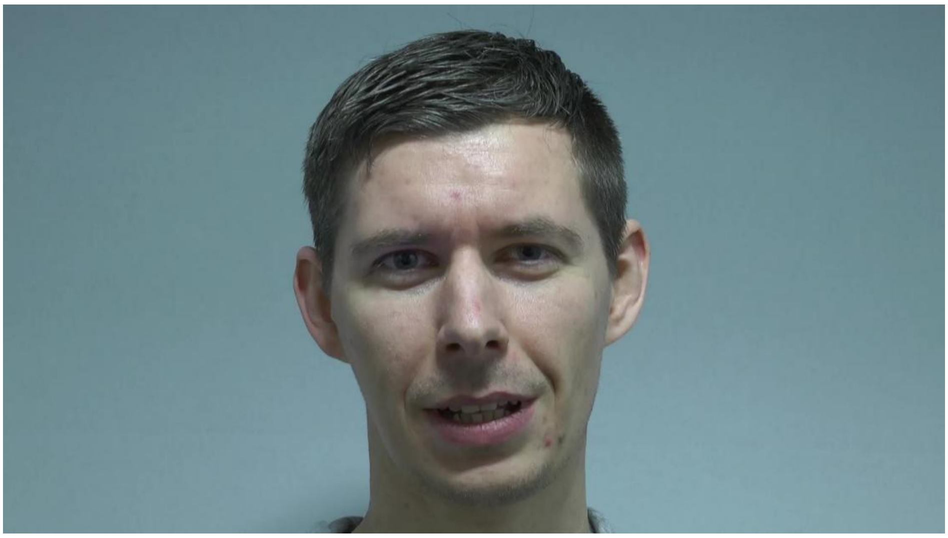


Figure 4.1 An example of the visual stimuli used in Experiment 1.

4.3.3.3 Background Stimuli

Background noise for use in the experiment was created using a database of talkers (Markham & Hazan, 2002) and was informed by the procedures outlined in Rosen et al. (2013). Sixteen male voices were chosen as sounding most similar to the target stimuli. Each talker was describing in free form language the scene they had witnessed on a video. This free form description ensured the talkers were not repeating the same text so little repetition of words was noticeable. Silences of more than 100ms were removed from each track and all filler expressions (e.g. erm, eh) removed to produce a continuous babble for each talker.

This process produced a thirty second track for each of the sixteen talkers. The talker chosen for the one talker background noise track was matched to the target talker as having a similar voice quality. Similarly, the talkers chosen for the two talker background noise track were the same talker used in the one talker track and a second similar sounding voice. The sixteen talkers processed were used for the sixteen talkers background track. Each of the tracks of the sixteen talkers was started at a different point along the main track to produce a 30 second track starting at different points along the story line to avoid any further repetition of words. A random 3 second section of the background noise was played with the target sentence in each trial.

4.3.3.4 Speech Processing

Sine-wave vocoded speech was used throughout this project for the audio and audio-visual stimuli. The stimuli were processed using Matlab software (Mathworks, Nantick, US). For each trial, the signal-to-noise ratio (SNR) was achieved by attenuating the target talker (for negative SNRs) or the background noise (for positive SNRs) before summing the two signals and normalising the root-mean-square of the composite signal. The composite signal was then band-pass filtered into eight frequency bands. The frequency bands were spaced equally on a bandwidth scale between 100Hz and 8kHz (Glasberg and Moore, 1990). The eight band-pass filters were summed for the clear speech conditions. For the vocoded speech conditions, the temporal envelope of each filter was extracted using the Hilbert transformation and used to modulate a sine wave at the centre frequency. The eight sine waves were summed to form the sine-wave vocoded speech stimuli. Eight bands were used as research has suggested that speech recognition performance does not improve if more than eight channels of information are available (Dorman et al., 1997).

4.3.4 Procedure

Participants were seated at approximately 0.5m from the display monitor with the display at head height. They were instructed to watch a video and listen to the audio stimuli (AV condition) or listen to the audio only (AO condition) and repeat any words they were able to understand at the end of each sentence.

4.3.4.1 Practice Trials

Participants undertook a practice session using E-Prime software in order to gain familiarity with the task. Four conditions were presented to participants, with five sentences in each condition; audio only clear speech, audio only vocoded speech, audio-visual clear speech, audio-visual vocoded speech. They were presented at a range of SNR’s from -8dB to 8dB in each type of background noise; one, two or sixteen talkers. A further three sentences were presented as visual only stimuli.

4.3.4.2 Main Experiment

For the main experiment, each participant was presented with thirteen conditions in a counter balanced order; audio only conditions with one, two or sixteen background talkers, and audio-visual conditions with one, two or sixteen background talkers for both clear and vocoded speech, and a visual only trial. Each condition was assigned a different list of twenty sentences from the 260 sentences used in a random order. Therefore, each participant had different sentences for each condition and in a different presentation order.

The first sentence in each condition was initially played to the participant at a SNR of -16db and the sentence repeated until three out of the five words were repeated correctly. Once the key words were correctly identified, a different sentence was presented on each subsequent trial. Once three out of the five key words were correctly identified in the initial sentence, the SNR was increased by 4dB until two reversals were made and then the reversals were reduced to 2dB for the remaining sentences. The SNR was increased by 2dB after a sentence was correctly identified and reduced by 2dB if a sentence was incorrectly identified. This resulted in an adaptive track for each condition for each participant. An example of an adaptive track is shown in Figure 4.2. The SNRs for the final ten sentences were averaged to produce a Speech Reception Threshold (SRT_{50}) for each participant, which represented the signal-to-noise ratio measured in decibels (dB) at which participants could report the key words within 50% of sentences correctly. This produced an average performance in dB for each condition for each participant. In the visual-only condition, participants were asked to verbally repeat any words they could understand at the end of each sentence and performance was scored in terms of percentage key words reported correctly.

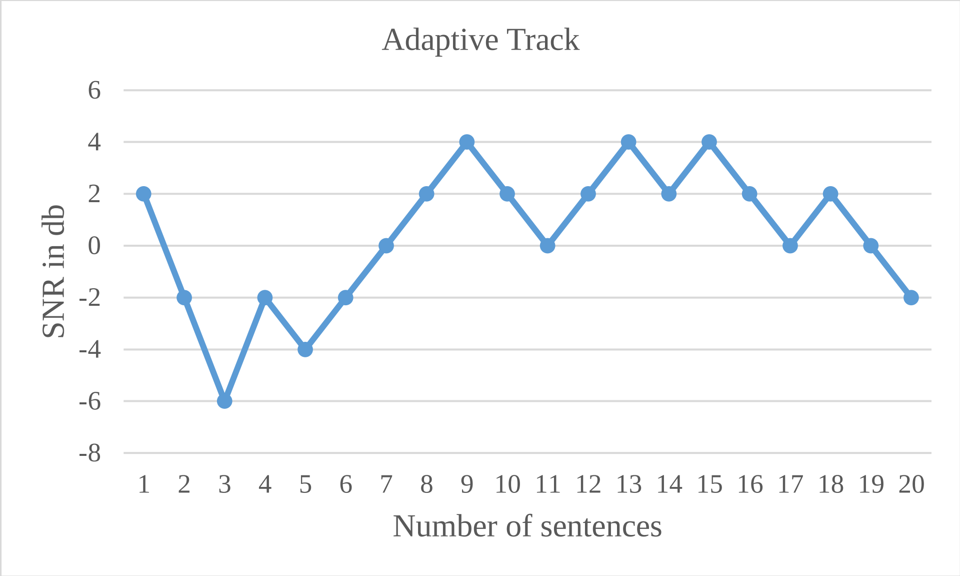


Figure 4.2 Example of an adaptive track obtained for each condition for each participant.

4.4 Results

4.4.1 Visual Only Results

The average visual only score was 1.52% correct (SD=2.7).

4.4.2 Audio Only and Audio-Visual Performance

SRT_{50s} were calculated for each condition for each participant. These were calculated by averaging the last ten trials on each participants' adaptive track for each condition, to give a mean SRT in dB. Figure 4.3 shows the mean performance score in dB for each of the conditions, audio only (AO) clear and vocoded speech, audio-visual (AV) clear and vocoded speech and for one, two and sixteen background talkers. For clear speech, average performance levels across all background noise types were 2.2dB for the audio only conditions (range 0.6dB to 3.3dB), and 0.2dB for the audio-visual conditions (range -3.4dB to 2.6dB). For vocoded speech, average performance levels were 12dB in the audio only conditions (range 10.7dB to 12.2dB) and 11dB in the audio-visual conditions (range 10dB to 11.4dB). There was considerable variation in performance between participants, as shown by the size of the whiskers.

A 2 (modality: audio only vs audio-visual) x 2 (speech type: clear vs vocoded) x 3 (background noise type, one, two and sixteen talkers) ANOVA showed a significant effect of modality ($F(1,22)=6.22$, $p=.021$, $\eta_p^2 = 0.22$). Performance was better in the audio-visual conditions than the audio only conditions. A significant main effect of speech type was also found ($F(1,22)=285.17$, $p<0.001$, $\eta_p^2 =.93$). Overall, clear speech was easier to understand than vocoded speech. Additionally, there was a main effect of the number of the background noise type ($F(2,44)=11.28$, $p<0.001$, $\eta_p^2 =.34$). Speech was easier to understand with one background talker than with two or sixteen talkers (Bonferroni corrected pairwise comparisons, one talker vs two talkers p=.003; one talker vs sixteen talkers p=.14; two talkers vs sixteen talkers p=.249). There was no significant interaction effect between modality, speech type and background noise type ($F(2,44)=2.44$, $p=.10$).

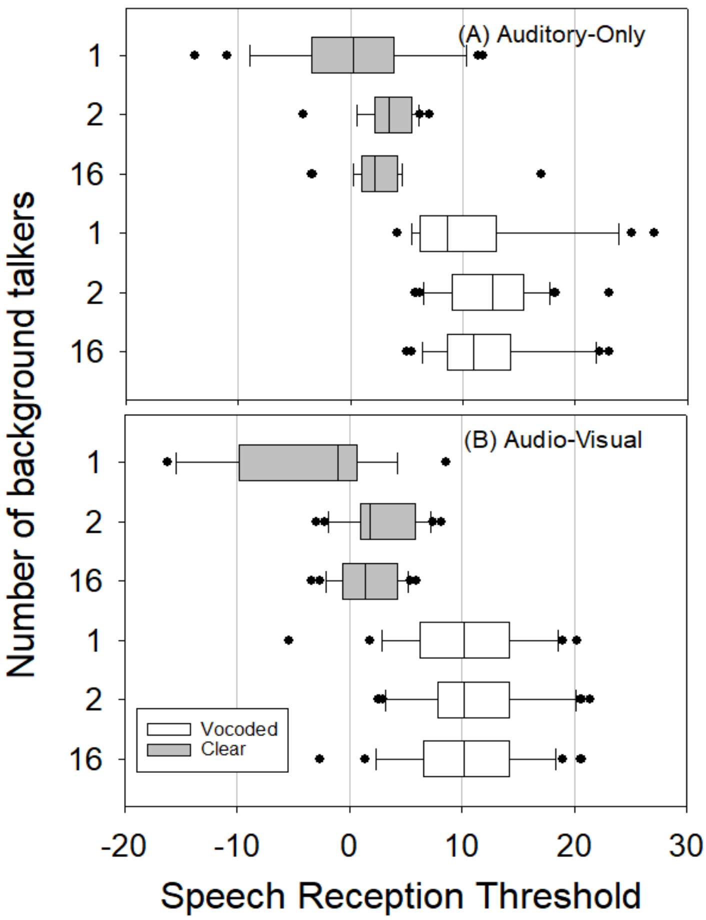


Figure 4.3 Average audio only (Panel A) and audio-visual performance (Panel B) for clear and vocoded speech. The rectangular boxes show the lower (25%) and upper (75%) quartiles of the data, with the solid line showing the median. The whiskers show the 10-90% range and the black dots show outlier data which fall outside that range.

4.4.3 Visual Speech Benefit

The amount of visual speech benefit received was calculated by measuring the difference between the AV and AO conditions. Figure 4.4 shows the average visual speech benefit for clear and vocoded speech, when there were one, two, or sixteen background talkers. There was considerable variation in the amount of visual speech benefit received by participants as shown by the size of the error bars.

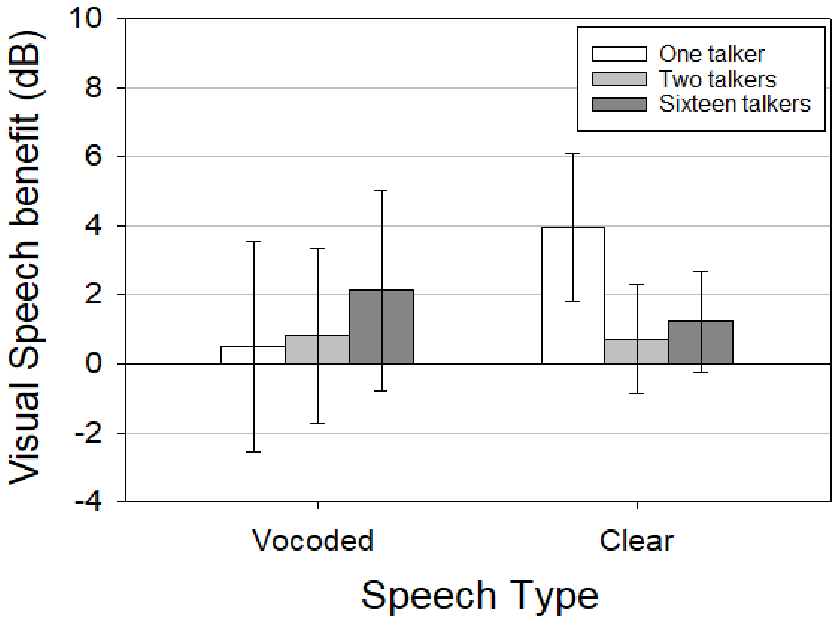


Figure 4.4. The amount of visual speech benefit received for clear and vocoded speech. Error bars denote 95% confidence intervals.

4.5 Discussion

Results show that as expected, clear speech was easier to understand than vocoded speech in each condition. When averaged across all conditions, significantly more speech could be understood if audio-visual information was presented rather than audio information alone. It was also hypothesised that increased visual benefit would be found for the vocoded speech over the clear speech, however, there was no significant difference in the amount of visual speech benefit in the vocoded conditions over the clear condition. The results also show that the size of the benefit received from the visual information was not as great as expected. Middelweerd and Plomp (1987) found an average of 4.6dB benefit to speech understanding with the addition of visual information when listening in background noise for younger listeners. In Experiment 1 the amount of benefit ranged from 0.5dB (vocoded speech with one background talker) to a maximum of 4dB (clear speech with one background talker) and with an average of 2dB benefit in clear speech and 1.2dB in vocoded speech across all types of background noise.

In addition, the SRTs whereby participants were able to identify the key words in the sentences were higher than expected. For example, previous research (Stacey et al, 2016) suggests that in a multi-talker babble, clear speech in an audio-visual paradigm can be recognised 50% of the time at an SNR of -8.8db. In Experiment 1, the average SNR for achieving a 50% correct response rate in the sixteen talker background noise in clear speech (similar to a multi-talker babble) was 1.5db. The addition of the visual information was only of significant benefit in the one-talker clear speech condition. Arguably, this condition was the easiest for participants; there were less background talkers and therefore more opportunities for glimpsing, clear speech is easier to understand than vocoded speech. Additionally, Rosen et al. (2013) found audio only performance for clear speech in noise-vocoded background noise was approximately 6-7dB’s in all conditions; one two and sixteen talkers. In Experiment 1, performance in audio only vocoded speech ranged between 10.7dB and 12.2dB, suggesting the task was more challenging than that conducted by Rosen et al. (2013).

Furthermore, six participants failed the adaptive track procedure for the audio only vocoded speech condition. This suggests that these participants found the experiment too challenging to complete in full, possibly because floor effects were taking place. There was also significant variation between participants in both the audio and audio-visual tasks and the amounts of visual speech benefit received, as evidenced the width of the whiskers and error bars (Figures 4.3 and 4.4). Variation in performance is greater in the vocoded speech conditions than the clear speech conditions; the whiskers and error bars are wider in the vocoded speech conditions than in the clear speech conditions. This supports the research of Rosen et al. (2013), who also found greater variation for clear speech in vocoded background noise than in clear background noise.

Taken together, this highlights the difficulty of this particular stimuli in this type of background noise. In this experimental paradigm, presenting visual information as an aid to speech perception was not enough to overcome the difficulty of the task in all of the conditions. The additional visual information was not enough to overcome the challenge of identifying the target voice from the background noise and therefore participants were unable to report the target voice, resulting in floor effects and large variation between participants.

Overall, the results of Experiment 1 suggest the difficulty of the task was too challenging for some participants and produced a large variability in performance in the remaining participants. Results in terms of the SRT's were lower than expected. Therefore, to enable participants to more successfully identify the target voice, additional cues may need to be provided in order to locate and report the target voice. To this end, Experiment 2 incorporated spatial cues in the form of an interaural delay.

4.6 Experiment 2

4.6.1 Introduction

The results of Experiment 1 demonstrate the difficulties in identifying the target voice in the background noise, producing floor effects and large variation in performance between participants. Therefore, to assist participants in locating the target voice, Experiment 2 repeated the methods of Experiment 1 with a change to the target stimuli. An interaural delay was introduced to the target voice (Carlile, 1996).

Research has shown that a release from masking takes place when spatial separation is achieved between the target voice and the masker (Kidd Jr, Mason, Deliwala, Woods & Colburn, 1994), therefore, it could be expected that introducing an interaural delay to the target stimuli would improve and stabilise performance. The auditory stimuli would be presented to participants from a different spatial location (left or right ear) to the background noise, enabling them to identity and report the target voice more easily. This may address the difficulties some participants had completing the task and reduce some of the large variability between participants. This would then produce a level of performance in line with other research and may change the amount of visual speech benefits participants receive.

The hypotheses for Experiment 2 were identical to those of Experiment 1;

- Clear speech would be easier to understand than vocoded speech
- Audio-visual performance would be better than audio only performance
- For clear speech and for vocoded speech, audio only and audio-visual speech perception performance would decline with an increase in the number of background talkers
- Greater benefit from visual speech information as the number of background talkers increased
- Greater levels of visual speech benefit would be found in vocoded speech than in clear speech.

4.7 Method

4.7.1 Design

The design of Experiment 2 was identical to Experiment 1.

4.7.2 Participants

Twenty-four participants took part in this experiment (age 19-47, mean age 28, 6 males). Participants were recruited from the student and staff population at Nottingham Trent University. Students were rewarded with research credits. Consent was obtained from each participant as agreed by the Nottingham Trent University Research Ethics Committee. Participants also confirmed normal hearing, normal or corrected to normal vision and had English as their first language. One participant was excluded for having incomplete data, leaving 23 participants data for analysis.

4.7.3 Materials and Procedure

The materials and procedure used in Experiment 2 were identical to those used in Experiment 1, with the following exceptions. Firstly, the experiments were conducted in a multi person IAC Acoustics 40a-5 audiology booth situated at Nottingham Trent University. Secondly, the target stimuli were presented with an interaural delay of 0.001 seconds. This was achieved by changes to the stimuli presentation using Matlab programming (Mathworks, Nantick, US). The target sentence was presented with a delay of 0.001 seconds in one ear, in advance of presentation to the other ear. The background noise was presented to both ears simultaneously. This resulted in a different time of arrival for the target sentence in each ear, providing location information to assist with identification of the target speech. The interaural delay was presented randomly to either the left or right ear throughout the experiment.

4.8 Results

4.8.1 Visual Only Performance

The average visual only score was 1.57% correct (SD=2.17).

4.8.2 Audio and Audio-Visual Performance

Audio only (AO) and audio-visual (AV) performance in dB was calculated in the same way as Experiment 1. All of the adaptive tracks were successful, with performance being between 30% and 70% correct as expected. Figure 4.5 shows the mean performance in dB for each of the conditions, audio only (AO) in clear and vocoded speech, audio-visual (AV) clear and vocoded speech and for one, two and sixteen background talkers.

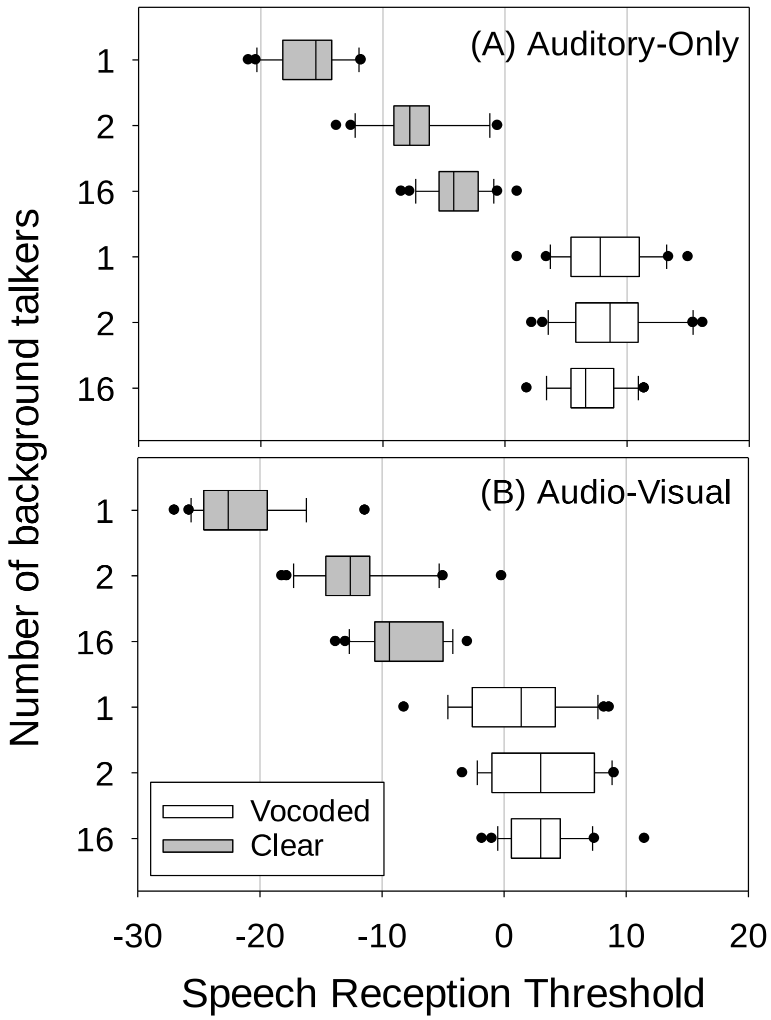


Figure 4.5 Average audio only (Panel A) and audio-visual performance (Panel B) for clear and vocoded speech. The rectangular boxes show the lower (25%) and upper (75%) quartiles of the data, with the solid line showing the median. The whiskers show the 10-90% range and the black dots show outlier data which fall outside that range.

Performance levels in Experiment 2 for the audio only and audio-visual conditions were significantly better than performance levels in Experiment 1. For example, the AO performance in the one talker clear condition in Experiment 1 averaged 0.6dB, the mean performance in Experiment 2 for this condition was -15.9dB. The average performance levels for clear speech across all background noise types were -9.2dB for the audio only conditions (range -15.9dB to -4dB), and -14.1dB for the audio-visual conditions (range -21.6dB to -8.6dB). For vocoded speech, average performance levels were 7.9dB in the audio only conditions (range 6.9dB to 8.7dB) and 2.4dB in the audio-visual conditions (range 0.9dB to 3.3dB).

A 2 (modality: audio only vs audio-visual) x 2 (speech type: clear vs vocoded) x 3 (background noise type, one, two and sixteen talkers) ANOVA showed a significant effect of modality ($F(1,22)=184.7$, $p=<.001$, $\eta_p^2 = 0.89$). Audio-visual speech was significantly easier to understand than audio only speech. A significant main effect of speech type was also found ($F(1,22)=881.14$, $p<.001$, $\eta_p^2 =.98$). Clear speech was easier to understand than vocoded speech. Additionally, there was a main effect of the number of the background talkers ($F(2,44)=161.01$, $p<.001$, $\eta_p^2 =.88$), with speech understanding being more challenging as the number of background talkers increased. This effect was mediated by an interaction with speech type (speech type x background noise type, $F(2,44)=84.11$, $p<.001$; $\eta_p^2 = 0.79$). Performance in clear speech was much more affected by the number of background talkers than performance in vocoded speech. Performance significantly decreased as the number of talkers increased for clear speech ($F(2,44)=221.05$, $p<0.001$, $\eta_p^2 = 0.91$), but for vocoded speech there was only just a significant difference between the different background noise types ($F(2,44)=3.34$, $p=.045$, partial eta squared = 0.13) and performance was worst in the two talker background noise.

4.8.3 Visual Speech Benefit

The amount of visual speech benefit received was calculated by measuring the difference between the AV and AO conditions. Figure 4.6 shows the average visual speech benefit for clear and vocoded speech, with one, two, or sixteen background talkers. The amount of visual speech benefit overall for vocoded speech was 5.5dB (range 3.8dB to 7.3dB), and for clear speech was 4.9dB (range 4.4dB to 5.7dB).

A 2 (speech type: clear vs vocoded) x 3 (background noise type, one, two and sixteen talkers) ANOVA showed there was no significant effect of speech type ($F(1,22)=.34$, $p=.57$, $\eta_p^2 =.015$). There was no significant difference between the amount of visual speech benefit between clear and vocoded speech. However, there was a significant effect of the background noise type ($F(2,44)=4.12$, $p=.022$, $\eta_p^2 =.16$). There was significantly more visual speech benefit in the one talker background noise than the sixteen talker background noise (Bonferroni corrected pairwise comparisons, one talker vs two talkers $p=.20$; one talker vs sixteen talkers p=.021; two talkers vs sixteen talkers p=1.0). There was no significant interaction between speech type and background noise ($F(2,44)=1.17$, $p=.32$, $\eta_p^2=.051$).

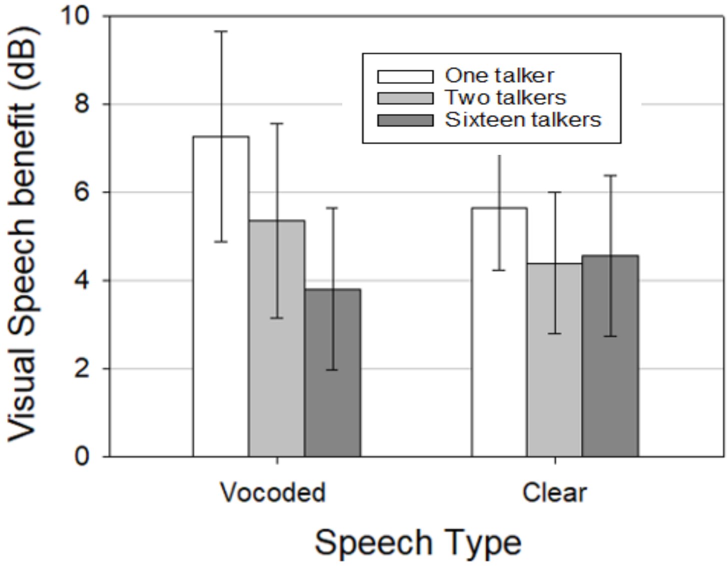


Figure 4.6 The amount of visual speech benefit received for clear and vocoded speech. Error bars denote 95% confidence intervals.

4.9 Discussion

4.9.1. Audio only and Audio-Visual Performance

As expected, the results of Experiment 2 showed that audio-visual speech was easier to understand than audio only speech, and clear speech was easier to understand than vocoded speech. Additionally, speech understanding was more challenging as the number of background talkers increased. Performance in clear speech was much more affected by the number of background talkers than performance in vocoded speech. Performance significantly decreased as the number of talkers increased for clear speech but for vocoded speech there was only just a significant difference between the different background noise types, with performance being poorest in the two talker background noise.

The addition of the interaural delay to the target stimuli significantly improved performance across all conditions in Experiment 2 compared to Experiment 1. For example, for clear speech, in audio only condition with one background talker, the average performance in Experiment 1 was 0.6dB, in Experiment 2 this increased to -16dB. This demonstrates a significant release from masking by the inclusion of the interaural delay. The overall amount of benefit received is also in line with previous research. Middelweerd & Plomp (1987) found an average gain of 4.6dB when visual information was presented alongside audio information in background noise, over audio information alone. Audio-visual gains in this study ranged from 4.4dB to 5.7dB for clear speech.

As hypothesised, the results of Experiment 2 showed that clear speech was easier to understand than vocoded speech. This was shown for every audio only and audio-visual condition, demonstrating the robustness of this effect. In addition, more speech could be understood if audio-visual information was presented rather than auditory information alone. This in contrast to Experiment 1, where there was no significant interaction effect between modality, speech type and background noise type. The visual information in Experiment 1 was not enough to provide benefit in the other conditions, despite the benefits that visual information is known to provide in speech understanding (e.g. Middelweerd & Plomp,1987). This suggests the inclusion of the interaural delay in Experiment 2 raised the threshold at which the visual information was of benefit to participants. The interaural delay was designed to improve audio only performance and allow listeners to identify the target voice from the background noise more easily. Once this was achieved, participants were able to successfully use the visual information to improve speech understanding. Overall, this suggests that visual information is only of benefit to participants once a certain audio threshold of understanding is achieved. If the audio information is particularly degraded, for example, by a particularly unintelligible talker or where the background noise is particularly challenging, the extra visual information is not enough to aid speech perception. It is possible that the talker in Experiment 1 may be less intelligible than other target talkers. This is discussed in detail in Chapter 5, where comparison of the intelligibly of target talkers is examined, and again in Chapter 8 where comparisons across experiments in this project are made.

The suggestion that visual information is only of benefit to participants once a certain audio threshold of understanding is achieved is also supported by the results of the vocoded speech conditions in Experiment 2. Performance in audio-visual conditions for vocoded speech was not affected by the inclusion of more background talkers. If the vocoded speech is particularly challenging for listeners, it could be expected that the addition of visual information would not improve performance as floor levels had already been achieved. This is in contrast to clear speech conditions where fewer background talkers led to improved audio-visual performance. The addition of the visual information in clear speech was enough to improve performance, with greater improvements with fewer background talkers.

It was also hypothesised that auditory-only performance would decline with the increase in background talkers due to increased masking effects (e.g. Brungart, 2001; Carhart, Johnson & Goodman, 1975; Cullington & Zeng, 2008; Humes, Kidd & Fogerty, 2017; Simpson & Cooke, 2005). For clear speech, audio only and audio-visual performance decreased with the addition of more talkers in line with this hypothesis. This suggests that for clear speech, as the number of talkers increase, masking effects increase, and speech perception becomes more difficult. This affects both the audio only performance and the audio-visual performance.

However, for vocoded speech, the auditory only performance was relatively unaffected by the addition of more background talkers. Therefore, masking effects may not be operating in the same way for vocoded speech as they are for clear speech. This is consistent with the results of Experiment 1. This suggests although the inclusion of the interaural delay improved performance across all conditions, it did not change the way masking effects operate in the vocoded speech conditions. Masking effects remain similar across the differing types of background noise. It could be argued the lack of TFS information in the vocoded speech increases masking effects. The lack of TFS information increases similarity between voices and voices are therefore less distinguishable from each other, meaning identification of the target voice becomes more challenging and the background talkers more similar. This would also increase the effects of informational masking across talkers, as the target voice would be more easily confused with the background talker or talkers.

Rosen et al. (2013) also found very little variation in auditory only performance for clear speech in a vocoded background noise of one, two, or sixteen talkers. Rosen et al. (2013) argue that the non-significant differences across vocoded background noise types reflects the lack of glimpsing opportunities in vocoded speech. There are few opportunities for glimpsing in vocoded speech and therefore floor effects are reached with very few background talkers.

The level of performance in Rosen et al. (2013) for clear speech audio targets in vocoded background noise was 6-7dB. In Experiment 2, the level of performance in the audio only vocoded speech conditions was between 6.9dB and 8.7dB, more in line with Rosen et al. (2013). The amount of visual speech benefit gained in Stacey et al. (2016) for vocoded speech in a multi-talker babble (similar to a sixteen talker background noise) was comparable to that gained in Experiment 2, 3.8dB. However, Experiment 2 included an interaural delay to the target stimuli to enable significant improvements to performance. An interaural delay was not used in the research of Rosen et al. (2013) or Stacey et al. (2016). This could suggest the talker used in Experiments 1 and 2 may be less intelligible than the talkers used in other research, or that the particular background noise used in Experiment 2, consisting of other talkers, led to lower SRT's in audio conditions. This is discussed in further detail in Chapter 5, where the relative intelligibility of talkers is compared and the effect on the amount of visual speech benefit gained by participants examined, and in Chapter 8, where comparisons of the intelligibility of the target talker used in Experiments 1 and 2 is made with other talkers in this project.

There was also a large variability in performance in the vocoded speech conditions. This was found in Experiment 2, as in Experiment 1; there was greater variability in the vocoded speech conditions as shown by the width of the whiskers and error bars, than in clear speech conditions (Figures 4.5 and 4.6). The addition of the interaural delay reduced variability from Experiment 1 to Experiment 2 (compare Figures 4.3 and 4.5), however, variability remains greater in vocoded speech than in clear speech. This may reflect a lack of experience with vocoded speech stimuli. Although participants undertook practice trials, research has demonstrated that adaptation to vocoded speech can be achieved with practice (Davis et al., 2005) it is therefore feasible these results may be stabilised with further training. Chapter 7 discusses familiarisation with vocoded stimuli and changes to performance levels as a result of increased experienced with vocoded speech materials.

4.9.2. Visual Speech Benefit

The level of visual speech benefit received was in line with previous research; for example, Middelweerd and Plomp (1987) reported a gain of 4.6dB in young normal-hearing listeners in clear speech. Experiment 2 showed a gain of between 4.3dB to

5.6dB for clear speech. This is in contrast to Experiment 1, where the amount of benefit ranged from 0.5dB to 3.95dB. This suggests the addition of the interaural delay allowed participants to identify the target audio stream more easily and gain more benefit from the accompanying visual information. This is discussed in more detail in section 4.9.1.

It was hypothesised that increased visual benefit would be found for the vocoded speech over the clear speech (Stacey et al., 2016), however, overall, no significant difference was found. In addition, for clear speech, there was no significant difference in the amount of visual speech benefit if the number of talkers in the background increased. For vocoded speech there was a close to significant difference between the one and sixteen talker noise types, with significantly more benefit in the one talker background noise than the sixteen talker background noise. This is in contrast to expectations, whereby where more benefit was expected when the signal is more degraded i.e. in vocoded speech over clear speech and in the sixteen talker noise over the one talker noise. This suggests the Principle of Inverse Effectiveness (Pof IE) does not apply universally to the amount of visual speech benefit gained as expected. The Pof IE asserts that where the audio stream of information is most degraded, additional information provided by the visual information stream may be at its most beneficial; as unimodal performance declines multi-sensory integration is improved (Meredith and Stein, 1986).

This supports the research of Tye-Murray et al. (2010) who argue the PofIE may not extend to visual speech perception and that integration of audio and visual stimuli for speech recognition may differ in the way they are integrated compared to other bimodal stimuli. Research has highlighted limitations to the amount of benefit that can be derived from visual speech information. The maximal point of visual speech benefit for clear speech were argued to be -12dB in pink noise (Ross et al., 2006) or -18dB in white noise (Altieri & Wenger, 2013). Overall, these differing results suggest there are limitations to the PofIE for speech perception, the maximal point dependent on the type of stimuli that is used and the type of background noise used and how it is measured.

Experiments 1 and 2 used IEEE sentences and background talkers. IEEE sentences are less predictable than other speech materials and may be more challenging for listeners than single word lists used in other research (Altieri and Wenger, 2013; Ross et al., 2006; Sumby & Pollack, 1954). There are also known increased masking effects from speech than other background noise types, such as pink or white noise (e.g. Brungart, 2001), which may change the amount of visual speech benefit that listeners are able to gain. The level of benefit was measured using an adaptive track procedure in Experiment 2 and not using a fixed performance level as used by other researchers (e.g. Altieri and Wenger, 2013; Ross et al., 2006). Further research could examine the stimuli used in Experiments 1 and 2 at fixed performance levels to establish at which SNR the maximum amount of visual speech benefit is gained. This would establish whether different stimuli have different maximum levels of visual speech benefit at differing SNRs. It could be argued that lower SNR's would be needed for speech stimuli because of the challenges of understanding speech due to increased masking effects. Additionally, if fixed performance levels were assessed and compared using the stimuli from Experiment 1 and Experiment 2 (with and without the interaural delay), it would establish whether the introduction of the interaural delay enabled listeners to gain more benefit from the visual speech information. If a difference between the SNR's at which the maximum visual speech benefit is gained is established, this would suggest the point at which the maximum visual speech benefit can be obtained is dependent on the audio intelligibly of the target talker. The relative audio intelligibility of the target talker is discussed in further detail in Chapter 5, where variability in the amount of visual speech benefit received is examined if the target talker changes.

4.10 Chapter Summary

- Experiment 1 examined the variability in the amount of visual speech benefit gained if different types of background noise were used in the test stimuli; one talker, two talkers and sixteen talkers, for clear speech and for vocoded speech
- Although a large variation in performance was found between individuals, no difference in the amount of visual speech benefit received was found between the background noise types or between clear and vocoded speech
- Performance levels were not as high as expected, based on previous research, suggesting participants were experiencing difficulty identifying the target speech from the background noise

- Experiment 2 included an inter aural delay to allow participants to identify the target voice more easily and achieve better performance levels
- In Experiment 2, audio only and audio-visual performance declined with the introduction of more background talkers in clear speech, however for vocoded speech, the increase in the number of background talkers did not change performance levels
- Masking effects are therefore argued to be operating more effectively in vocoded speech, resulting in little opportunities for glimpsing and increased masking, resulting in floor effects with very few background talkers
- The amount of visual speech benefit gained did not vary significantly as the type of background noise changed for clear speech or for vocoded speech, and was not greater in vocoded speech than in clear speech as expected
- This suggests the Principle of Inverse Effectiveness, whereby multi-sensory integration is enhanced as unimodal performance declines, may not apply monotonically to visual speech perception tasks.

Chapter 5 – Changes to the Target Talker

Chapter 5 reports the third experiment in this project; changing the target talker in the test stimuli. It could be argued, that as talkers demonstrate different levels of audio intelligibility and different levels of visual only benefit, differences in the levels of visual speech benefit that talkers provide could be expected. In order to assess this, Experiment 3 includes four different talkers who have different levels of auditory-only intelligibility. Variability in performance levels and changes to the amount of visual speech benefit gained by participants is assessed for both clear and vocoded speech.

5.1. Introduction

This chapter examines changes to the amount of visual speech benefit received by participants if the target talker changes. Chapter 4 examined variation in the level of visual speech benefit received if the background noise changes. In Experiment 3, examined in Chapter 5, the stimuli is manipulated in a different way; the target talker varies but the background noise remains the same.

Experiments 1 and 2 demonstrated how levels of visual speech benefit vary as a result of changes to the type of background noise. Using a single target talker in Experiments 1 and 2 eliminated the costs of switching between talkers. In Experiment 1, it could be argued that the talker used in the stimuli was particularly unintelligible and the background noise type especially demanding, and therefore expected performance levels in terms of SNR's were not achieved. To address this, an interaural delay was added to the target talker and performance was subsequently improved to expected levels in Experiment 2.

It is possible that the audio intelligibility of the target talker may change performance levels in audio only and audio-visual tasks and change the amount of visual speech benefit received by participants. There are known differences in audio intelligibly between talkers (e.g. Gagne, Masterson, Munhall, Bilida & Querengesser, 1994) but it is not known whether changing the audio intelligibility of the target talker changes the amount of visual speech benefit gained. It is feasible that the target talker used in Experiment 1 and 2 is of greater or less intelligibility than those used in other research (e.g. Altieri & Wenger, 2013; Rosen et al., 2013;

Ross et al., 2006; 2013, Stacey et al., 2016) and therefore differences in results may be attributed to the intelligibility of the talker or talkers used within the research stimuli.

In Experiment 3, the background noise is kept constant and the target voice varies. This addresses whether changes to the target talker that change audio and audio-visual performance levels, also change the amount of visual speech benefit gained. This is assessed for both clear and vocoded speech. Variation in the amount of visual speech benefit was found in Experiments 1 and 2 between clear speech and vocoded speech. It is not known if this variation was specific to the target talker used in Experiment 1 and 2, or if variation will be found for all talkers. Therefore, the amount of visual speech benefit may vary between clear and vocoded speech for different target talker.

5.2 Experiment 3 – Changing the Target Talker

5.2.1 Differences in the Intelligibility of Talkers

Despite known variability between talkers, understanding speech remains a relatively easy task to achieve. The ‘lack of invariance’ in speech perception (Liberman, Cooper, Shankweiler & Studdert-Kennedy, 1967) suggests that talkers physically produce speech in different ways. Despite this large variability across talkers, listeners are able to generalise across talkers and cope well with differences in speech production. However, it is also argued different talkers have different levels of intelligibility, which can make speech perception more challenging (Lesner, 1988).

5.2.1.1 Visual Information

Lesnar (1988) argues there are a variety of factors that contribute to the ease with which talkers can be understood. Research has found that target talkers vary in their visual only intelligibility (Kricos & Lesner, 1982; Lachs & Hernandez, 1998; Montgomery & Jackson, 1983). For example, using everyday sentences as the test stimuli, Demorest & Bernstein (1992) found 4.9% of the variance in performance in visual only tasks was accounted for by differences in the two target talkers used in the study. Similarly, Lachs and Hernandez (1998) found that intelligibly scores ranged between approximately 5% and 22% correct between ten target talkers using single words as the test stimuli.

Using visual only stimuli, Conrey and Gold (2006) found that different talkers have different physical characteristics in the way speech is expressed. They assessed participants using eight different target talkers and found that the target talkers showed variation in facial movements. Focusing on the mouth area as an aid to speech perception has been shown to be an effective strategy in speech perception, particularly in background noise (e.g. Buchan, Paré & Munhall, 2008). However, focusing on the mouth did not result in higher performance levels for some target talkers in Conrey and Gold (2008). This suggests talkers convey speech information in physically different ways. Conrey and Gold (2006) go on to argue that focussing on other facial areas, other than the mouth region, may be a more effective strategy for understanding some target talkers.

The difference in visual only intelligibility may contribute to the amount of visual speech benefit that target talkers are able to provide and is therefore an important consideration and is examined in Experiment 3.

5.2.1.2 Audio and Audio-Visual Information

Talkers have also been shown to differ in their audio intelligibility (Bond & Moore, 1994; Bradlow, Torretta, & Pisoni, 1996; Hazan & Markham, 2004). When single words were presented at 6dB in a 20 talker babble, Hazan & Markham (2004) found significant differences across 45 target talkers. When scored as percentage of the key words correctly identified, scores ranged between 81.2% and 96.4%. Similarly, Cox, Alexander and Gilmore (1987) tested six talkers using a multi-talker babble background noise in four different reverberation environments; a living room, a large classroom, a large social event room, a larger social event room. They found significant differences in intelligibility between the six talkers tested, with the most intelligible talker being the most intelligible across all listening environments. However, one of the talkers used had significantly reduced intelligibility in one of the environments, the large classroom. This environment had enhanced reverberation effects and the talker was further away from the listener. This suggests that not all voices are affected in the same way by degradation effects. Intelligibility of the talker may vary as a function of the background noise being tested.

Lachs and Hernandez (1998) reviewed the audio only, visual only and audio-visual stimuli held in the Hoosier Audiovisual Multi-Talker Database (Sheffert, Lachs & Hernandez (1996). The database consists of ten talkers, five male and five female, and 3000 single words. Participants were asked to correctly identify the words in the three presentation modes and were scored as percentages correct. Significant variation between the ten talkers was found in the audio only, visual only and audio-visual stimuli. In addition, some talkers had differing performance levels in different conditions. For example, one male talker was found to have scores of approximately 5% correct in the visual only condition, 85% correct in the audio only condition and 98% correct in the audio-visual condition; an overall visual speech benefit increase of 13%. The next male talker had a visual only score of 22% correct, and audio only score of 91% correct and an audio-visual score of 88% correct; an overall visual speech deficit of -3%. This demonstrates the large variation between talkers in all presentation modes, even within established research stimuli. It also suggests the same target talker may be more intelligible in one presentation mode than another mode, and the amount of visual speech benefit provided may not correlate with the audio only or visual only performance level.

Similarly, Gagne et al. (1994) assessed levels of intelligibility for 10 different target talkers in visual only, audio only and audio-visual conditions. Test stimuli consisted of single words and were degraded in the audio and audio-visual modality by simulating perceptual hearing loss to prevent ceiling effects. In order to remove the possibility of floor effects in the visual only condition, a pilot study was conducted to identify simple words that would result in test scores of approximately 60%. This resulted in relatively high scores in the visual only condition of between 61 and 84% correct. In the audio only condition, scores ranged between 10-40% correct, and in the audio-visual condition ranged between 56-77% correct. There were significant differences between talkers in each of the conditions, although statistical analysis between presentation modes was not performed by Gagne et al. (1994) as the study was carried out between participants. However, Gagne et al. (1994) go on to argue the visible difference between conditions suggests that target talkers who exhibit greater levels of intelligibility in one modality do not necessarily exhibit the same advantage in another modality. Although some target talkers granted a similar advantage across modalities, this did not uniformly apply across talkers. For example, Talker B was the second most intelligible in the audio only condition but the eighth most intelligible in the visual only condition. Conversely, Talker G was the most intelligible in the visual only condition but the seventh most intelligible in the audio only condition. Gagne et al. (1994), also argue the results in the audio-visual speech condition somewhat reflect the average score of the target talker across the audio only and audio-visual conditions. However, on examining the reported data, there does not appear to be a noticeable relationship between correct scores achieved in the audio only and visual only conditions and scores in the audio-visual condition. This would suggest that the benefit to speech understanding talkers are able to provide in audio-visual conditions does not have a monotonic relationship with performance levels in audio only and audio-visual conditions. Performance levels vary across talkers and between modalities for those talkers.

Research has also assessed difference in audio intelligibility using vocoded stimuli. Green, Katiri, Faulkner and Rosen (2007) found that for clear speech presented at 6dB in background noise types of a 20 talker babble, four channel and eight channel noise vocoded speech, the relative intelligibility of the six target talkers remained the same. The relative level of intelligibility for each taker did not vary if the speech was clear or vocoded, although significant differences in levels of intelligibility were found between the six target talkers. In a comparable study, Bent, Buchwald and Pisoni (2009), found a significant correlation between levels of intelligibility for target talkers between eight channel sine wave vocoded speech and clear speech in a six talker babble presented at 0dB.

Using an audio-visual paradigm, Lander and Davies (2008) assessed speech reading ability of participants using single words at -20dB. They found that a familiarisation task with a talker in an audio-visual presentation significantly improved speech reading (visual only) ability with the same talker. The same level of improvement was not found when participants gained familiarity with a different talker. This suggests that participants were gaining speaker specific information through exposure to the talker which aided understanding; different talkers have different characteristics. Only two talkers were assessed in Lander and Davies (2008) and no difference was observed between the talkers.

Other cross modal effects have been argued to take place during speech perception, suggesting speaker specific characteristics are observed by participants. Rosenblum, Miller and Sanchez (2007) found that performance in audio only speech perception is improved by prior access to visual only information from the same talker. They assessed visual only performance using BKB sentences (Benck & Bamford, 1979), followed by audio only performance levels in background noise, with the same talker or a different talker. Audio performance levels were better when the target talker had previously been viewed in the visual only condition. Rosenblum et al. (2007) argue cross modality effects are talking place to enable better speech perception in noise. Similarly, when participants had heard from a talker prior to trials in lip reading, visual only speech perception for those talkers improved (Sanchez, Dias & Rosenblum, 2013). This suggests that different target talkers have individual characteristics that are assimilated by participants to aid speech understanding from that talker.

Considering the known variability between talker intelligibility, it might be problematic that some researchers use different talkers within the same experiment (e.g. Brungart, 2001; Brungart & Simpson, 2007; Josupeit, & Hohmann, 2017). For example, Brungart (2001) used four female and four different male voices to assess differences in masking effects. Assessment of the audio intelligibility of the eight target talkers is not mentioned in the research. If the audio intelligibility of the target talkers varied, this may have changed the masking effects within the background noise, and therefore changed the results of the study.

Taken together, this suggests there are differences across talkers, in terms of the visual information and auditory information they are able to provide, that aid speech understanding. It is also worth noting that different participants were used to assess the performance between conditions by Lachs and Hernandez (1998) and Bent et al. (2009). Differences between participants may have changed performance outcomes. Differences between participants is discussed in detail in Chapters 6 and 7 of this project.

5.2.3 Principle of Inverse Effectiveness (PofIE)

As outlined in section 4.2.2, the Principle of Inverse Effectives (PofIE) (Meredith & Stein, 1986) suggests that as unimodal performance declines, multisensory integration is improved. It could therefore be expected that if the audio signal is degraded, due to lack of intelligibility from the target talker, the benefit from visual information would increase. The most unintelligible talkers may provide the greatest levels of visual speech benefit.

Experiment 1 and 2 in this project showed that the PofIE does not apply monotonically to speech perception where the number of talkers in the background noise varies. It is not known whether the PofIE similarly applies to speech perception when the target talker changes in intelligibility. This is therefore assessed in Experiment 3.

5.2.4 Summary

It is not known whether the changes to the visual only and audio only performance, outlined in sections 5.2 and 5.3, change the level of visual speech benefit participants receive. It could be argued, however, that as talkers demonstrate different levels of audio intelligibility, differences in the levels of visual speech benefit that talkers provide could be expected. In order to assess this, Experiment 3 included four different talkers who had different levels of auditory-only intelligibility. The levels of visual speech benefit were measured for these four talkers for both clear and vocoded speech.

In addition, if the Principle of Inverse Effectiveness (PofIE) were to apply across talkers, it would be expected that the talkers with the lowest audio intelligibility would provide the most visual speech benefit. Similarly, if the speech is degraded by vocoding, more visual speech benefit would be found. Overall, therefore, it could be expected that the most visual speech benefit would be found for the least intelligible talker in the vocoded speech condition and the least visual speech benefit for the most intelligible talker in the clear speech condition. It has also been argued that talkers in vocoded speech and clear speech in noise have similar intelligibility (Bent et al. 2009). Therefore, changes in performance levels between clear and vocoded speech may be attributed to differences in the multisensory integration of the audio and visual streams of information and the relative visual benefits of the target talker, and not attributed to differences in audio intelligibility between clear and vocoded speech for the target talker.

However, the results of Experiments 1 and 2 showed that for the particular target talker used, there was no difference in the levels of visual speech benefit between clear and vocoded speech. This is in contrast to other research (Stacey et al., 2016). It is possible this is a result of the characteristics of the target talker used in Experiments 1 and 2. If a different target talker was used, more benefit from vocoded speech conditions may be found than in clear speech conditions. This is examined in Experiment 3; four different target talkers are assessed for both clear and vocoded speech.

Lesner (1988) also argues there are a number of modifiable traits that can make speech easier to understand. These include; a slower speech rate, precise articulation of words, additional pauses in speech and appropriate gestures. The target talkers used in Experiment 3 were asked to speak at their usual rate and with their usual articulation, to control for any additional emphasis that may increase speech understanding. The sentences used were IEEE (Rothauser et al., 1969) which do not have much opportunity for including pauses. Additionally, IEEE sentences are open set and have very few contextual cues. Therefore, performance would be focused on the articulatory aspects of the speech and not linguistic information. The visual stimuli consisted of only the head region for each talker to remove any possible benefit from gesticulation.

5.2.5 Hypotheses for Experiment 3

It was hypothesised that clear speech would be easier to understand than vocoded speech, and that audio-visual performance would be better than audio only performance.

For clear speech and for vocoded speech it was hypothesised that greater levels of visual speech benefits will be found for less intelligible talkers.

It was also hypothesised there would be more visual speech benefit in vocoded speech than in clear speech.

5.3 Method

The methods were the same as Experiment 2 with the following exceptions.

5.3.1 Design

A within-participants design was used in this study to assess audio and audio-visual performance. Audio only (AO) and Audio-visual (AV) Speech Reception Thresholds (SRT_{50}), measured in decibels (dB) were obtained for clear and vocoded speech for each of the four target talkers.

The first analysis was a 2 (modality: audio only vs audio-visual) x 4 (target talker identity) ANOVA for clear speech. The dependent variable was SRT's measured in dB.

The next set of analyses examined the vocoded speech conditions. Talker 4 was excluded from the main analysis due to the failure of some of the adaptive tracks for this talker. This is discussed in detail in section 5.4.2. The remaining analysis was therefore conducted as follows;

The second analysis was a 2 (speech type: clear vs vocoded) x 3 (target talker identity) ANOVA for the vocoded speech. The dependent variable was SRT's measured in dB.

The third analysis was a 2 (speech type: clear vs vocoded) x 3 (target talker identity) ANOVA for the amount of visual speech benefit received.

A further analysis of the performance levels of Talker 4 in the vocoded speech conditions is reported.

A separate experiment measured performance in a visual only task for each of the four target talkers and was scored as percentage correct.

5.3.2 Participants

Twenty-four participants (age 18-33, mean 22, 2 male) took part in this experiment. All other criteria were repeated from Experiment 1.

5.3.3 Materials and Procedure

Pilot Study

In order to address the variation in performance using different target talkers, eleven different talkers were recorded articulating the same 30 IEEE sentences (Rothauser et al., 1969). The relative audio intelligibility of these talkers was assessed by conducting a pilot study using six participants. Participants were asked to identify key words in the 330 IEEE sentences. The sentences were presented in a random order in clear speech at an SNR of -8dB in a 16 talker background noise. The percentage of key words correctly identified was recorded for each of the eleven talkers.

Results showed a large variation in intelligibility between talkers, with overall mean correct scores ranging from 45% to 88% correct (Table 5.1). Four talkers were selected for use in the main experiment; the two talkers with the highest ranked intelligibility (one male, talker 9; and one female, talker 2) and the two talkers with the lowest ranked intelligibility (one male, talker 6; and one female, talker 3).

Table 5.1 Percentage of words correct identified for each of the 11 talkers in the auditory pilot study. Figures in bold denote talkers taken through to the main experiment.

Talker	Mean	SD
1	66.5	14
2	**82.2**	**12.7**
3	**44.7**	**17.3**
4	63.1	11
5	67.8	10.8
6	**54.2**	**19.7**
7	66.9	22
8	72.4	10.3
9	**88.4**	**7.2**
10	80.7	12
11	69.6	19.5

Practice Block

Participants were presented with a practice block before the main experiment using Matlab programming (Mathworks, Nantick, US). A different talker to those used in the main part of the experiment was used to present 20 IEEE sentences. Five sentences were presented for each of the conditions; audio and audio-visual in clear speech, audio and audio-visual in vocoded speech. The audio stimuli were presented at an initial SNR of -8dB, the target talker was then reduced in volume in 4db steps if 3 out of the 5 words were correctly identified or increased in 4dB steps if the key words were not identified.

Main Experiment

For the main experiment, the four target talkers recorded eighty different IEEE sentences each (Rothhauser et al., 1969). Therefore, 320 sentences were presented in total. The experimental methods used in Experiment 2 were repeated for these four talkers but were conducted using only one background noise type, the sixteen talker noise. A sixteen talker background noise type was created using both male and female voices from the UCL database (Markham & Hazan, 2002), and processed as outlined in Experiment 1. The sixteen talker background noise type was selected as being the closest to that used in previous research where a multi-talker babble was used (Stacey et al., 2016). In addition, the results of Experiment 2 showed that the greatest variation in performance was found in this noise type, and therefore greater variation in performance could be expected in Experiment 3.

Each talker and each condition was presented in a random order for each participant, the four conditions being; audio only in clear and vocoded speech, audio-visual in clear and vocoded speech for each of the four target talkers. The twenty sentences used in each condition were also counterbalanced from the 80 sentences available from each talker, so each participant had a different list of 20 sentences for each talker. These 20 sentences were presented in a random order. Therefore, each participant had different sentences for each talker and in a different presentation order. The first sentence in each condition was presented at an initial SNR of -16db and an adaptive track procedure followed, as outlined in Experiment 1, to produce an average SRT for each participant and for each condition. These were averaged to give an average overall performance for each of the four target talkers for each condition.

5.3.4 Visual Only Performance

Visual only performance was measured in a separate experiment. Twenty-four participants (age 19-31, mean 21.5, 7 male) took part in this experiment. Each participant was presented with 20 sentences from each of the four talkers. The 20 sentences were presented in a random order for each talker and the order of the talkers was counterbalanced for each participant. Participants were asked to verbally repeat any words they could understand at the end of each sentence and a percentage score correct was recorded.

5.4 Results

5.4.1 Visual Only Performance

Percentage scores correct for each talker are shown in Table 5.2. As this was an open-set test chance performance is 0%.

Table 5.2*:* Percentage of key words correctly identified for each talker in the visual only condition. Standard deviations are shown in brackets.

Talker	Percentage of key words correctly identified
1	2 (2.7)
2	1.8 (2.6)
3	1.4 (2.1)
4	0.6 (1.1)

5.4.2 Audio and Audio-Visual performance

Audio only (AO) and audio-visual (AV) performance in dB was calculated in the same way as Experiment 1. All adaptive tracks were successful for Talkers 1, 2 and 3, with performance being between 30% and 70% correct as expected. However, for Talker 4, in the audio only vocoded speech condition, seven of the participants had overall levels of performance of between 10 and 20%, indicating that the adaptive tracks had failed for this condition. Figure 5.1 shows an example of a failed adaptive track for one participant in the audio only condition for Talker 4. The average performance level across the 24 participants was 34.2% correct. Therefore, the resulting SRT's could not be reliably included in the analysis for vocoded speech. Performance for vocoded speech is therefore reported for the Talkers 1, 2 and 3 only. The results for vocoded speech for Talker 4 are reported separately in section 5.4.4. For clear speech, results for all four talkers are reported. Comparisons between performance levels in clear and vocoded speech are made between Talkers 1, 2 and 3 only.

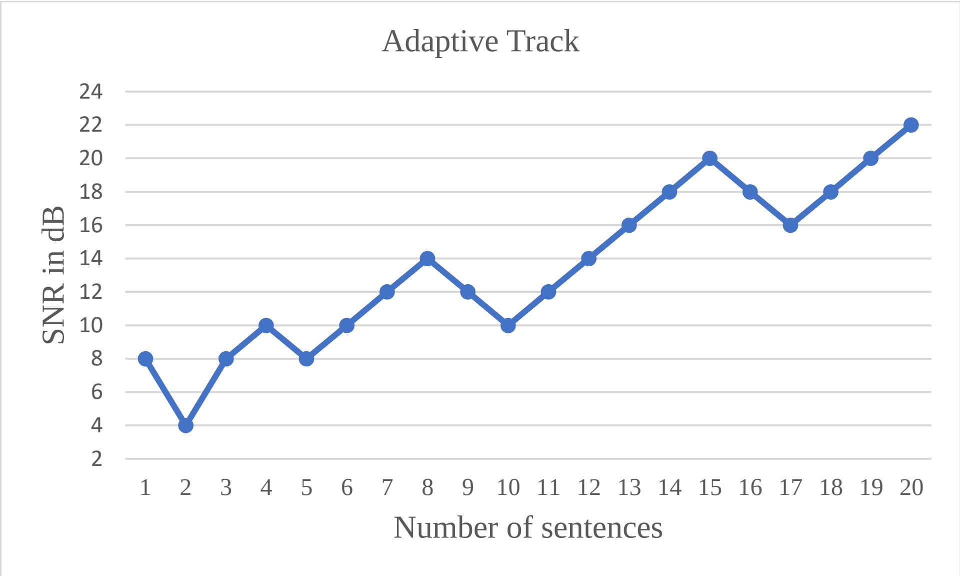


Figure 5.1 An example of a failed adaptive track for one participant in the audio only condition for Talker 4.

5.4.2.1 Clear Speech

Figure 5.2 shows the mean performance levels in dB for audio only (AO) and audio-visual (AV) for clear speech for each of the four target talkers. Talkers are ordered according to their audio intelligibility and not the order of presentation to the participants.

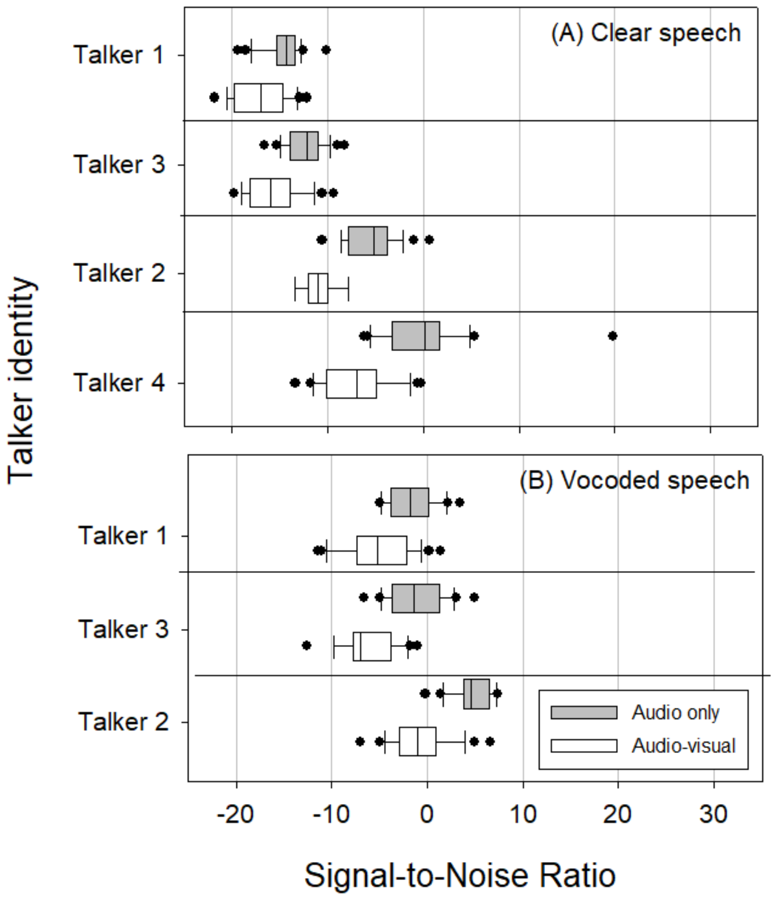


Figure 5.2. Audio only and audio-visual Speech Reception Thresholds for Clear (Panel A) and Vocoded (Panel B) speech for the four different talkers. Talkers have been ordered according to their intelligibility in the Audio only condition for clear speech, not the presentation order. Three talkers are shown for vocoded speech due to the failure of the adaptive tracks for Talker 4 in the vocoded speech condition. The whiskers show the 10-90% range and the black dots show outlier data which fall outside that range.

A 2 (modality: audio only vs audio-visual) x 4 (target talker identity) repeated measures ANOVA showed a significant effect of modality ($F(1,23)=167.28$, $p<.001$, $\eta_p^2 = .88$). Audio-visual speech was easier to understand than audio only speech. There was a significant effect of talker identity ($F(3,69)=251.77$, $p<.001$, $\eta_p^2 = .92$). In line with the results from the pilot study, audio-only performance was best with Talker 1 (mean = -14.67dB, SD=2.0) and much more favourable SNR's were needed to understand Talker 4 (mean = 0.25dB, SD=5.23). There was a significant interaction effect between modality and talker identity (($F(3,69)=11.17$, $p<.001$, η_p^2

=.33). This is examined in further detail in section 5.4.3.1, where the amount visual speech benefit gained is outlined.

Additionally, there was considerable variation between participants, as shown by the width of the whiskers in Figure 5.2. The variation was greatest for Talker 4, the most unintelligible talker.

5.4.2.2 Vocoded Speech

As outlined in section 5.4.2, results for the vocoded speech conditions include Talkers 1, 2 and 3 only. The average performance levels for these talkers are shown in Figure 5.2. Talkers are ordered according to their intelligibility in clear speech. The order of performance levels across the talkers in vocoded speech differed slightly to those in clear speech. Talker 1 was the most intelligible in clear speech but the second most intelligible overall in vocoded speech, with an average SRT in the audio only condition of -1.5dB (SD=2.33) and an average SRT of -5dB (SD=3.32) in the audio-visual speech condition. Talker 3 was the most intelligible overall, with average SRT's of -1.23dB (SD=3.03) in the audio only condition and -6.23dB (SD=2.82) in the audio-visual condition. Talker 2 was the least intelligible in vocoded speech (audio only; M=4.68dB, SD=2.01, audio-visual; M=-0.78dB, SD=3.05).

A 2 (modality: audio only vs audio-visual) x 3 (target talker identity) ANOVA showed a significant effect of modality ($F(1,23)=152.54$, $p<.001$, η_p^2 =.87). Audio-visual speech was easier to understand than audio only speech. There was a significant effect of talker identity ($F(2,46)=117.88$, $p<.001$, η_p^2 =.84), with Talker 2 being the least intelligible and Talker 1 the most intelligible. There was no significant interaction effect between modality and talker identity ($F(2,46)=2.21$, $p=.121$, η_p^2 =.09).

5.4.3 Visual Speech Benefit

The amount of visual speech benefit received was calculated by measuring the difference between the audio only and audio-visual conditions.

5.4.3.1 Clear Speech

The amount of visual speech benefit in clear speech is shown in Figure 5.3 (Panel A). This is shown for all four talkers. The Talkers are ordered according to their audio intelligibly in clear speech. There was significantly more visual speech benefit provided by Talker 4 than Talkers 1 and 3 ($t(23)=3.85$, $p=.006$; $t(23)=3.2$, $p=.024$). There was significantly more visual speech benefit provided by Talker 2 than Talkers 1 or 3 ($t(23)=5.4$, $p<.001$; $t(23)=3.5$, $p=.012$). There was no significant difference between the amount of visual speech benefit provided by Talkers 1 and 3 ($t(23)=-2.45$, $p=.132$), and between Talkers 2 and 4 ($t(23)=-1.79$, $p=.48$). Talker 1 had the highest level of audio intelligibility and the lowest amount of visual speech benefit in clear speech. Likewise, Talker 4 had the lowest intelligibility and the highest amount of visual speech benefit. The amount of visual speech benefit decreased uniformly with the audio and audio-visual intelligibility in clear speech.

In addition, the variability in the amount of visual speech benefit gained was significantly greater for Talker 4. This is shown by the width of the error bar for Talker 4 (Figure 5.3).

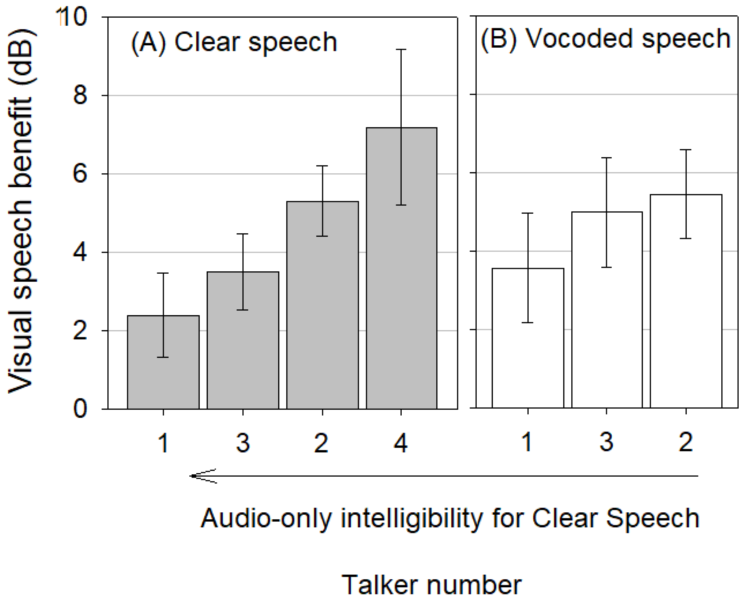


Figure 5.3 The amount of visual speech benefit for clear speech (Panel A) and vocoded speech (Panel B). Talkers are ordered according to intelligibility in clear speech. Error bars denote 95% confidence intervals. Three talkers are shown for vocoded speech due to the failure of the adaptive tracks for Talker 4 in the vocoded speech condition.

5.4.3.2 Vocoded Speech

The amount of visual speech benefit for vocoded speech is shown in Figure 5.3 (Panel B). This is shown for Talkers 1,2 and 3. The talkers are ordered according to their intelligibility in clear speech. There were no significant differences in the amount of visual speech benefit received between each talker, once adjusted for the Bonferroni correction. As found for clear speech, the amount of visual speech benefit increased as the intelligibility of the talker decreased in clear speech. However, the amount of benefit in vocoded speech did not correspond to the audio intelligibly in vocoded speech. Overall, the most intelligible talker in vocoded speech was Talker 3, but this talker did not have the lowest amount of visual speech benefit.

5.4.3.3 Comparison of Clear and Vocoded Speech

Comparison of the amount of visual speech benefit in clear and vocoded speech is made for Talkers 1, 2 and 3 only. Figure 5.3 shows the amount of visual speech benefit for clear and vocoded speech for Talkers 1, 2 and 3. As shown by the width of the error bars, there was more variation in the amount of visual speech benefit received in vocoded speech than in clear speech.

5.4.4 Vocoded Speech for Talker 4

As outlined in section 5.4.2, the adaptive tracks for seven participants failed for the audio only vocoded speech condition for Talker 4. The vocoded speech results for Talker 4 were therefore not included in the main analysis above. However, results for vocoded speech for the remaining 17 participants are presented in this section.

Audio only performance for Talker 4 was an average of 18.41dB (SD=7.03) and in the audio-visual condition was 6.76B (SD=3.20). The amount of visual speech benefit was on average of 11.65 (SD=7.55). Therefore, although direct comparison is not made with Talkers 1, 2 and 3, it can be seen that in line with clear speech, Talker 4 was the most unintelligible in the audio only and audio-visual conditions for vocoded speech. Additionally, the amount of visual speech benefit for Talker 4 was greater than the other talkers. Therefore, the amount of visual speech benefit for vocoded speech increased uniformly in line with talker intelligibility for Talker 4, as found in the clear speech conditions.

5.5 Discussion

As expected, variation in performance levels were found between the four target talkers. Additionally, more visual speech benefit was found in vocoded speech conditions than in clear speech conditions overall. The amount of visual speech benefit received varied as a result of the audio intelligibility of the target talker and whether the speech was clear or vocoded. For clear speech, the amount of visual speech benefit decreased as the audio and audio-visual intelligibility increased. These results are discussed in detail in the next section.

5.5.1 Audio and Audio-Visual Performance

As expected, speech perception performance was better when speech was clear then when it was vocoded. Performance was also better for audio-visual conditions than audio only conditions. In line with Experiment 1 and 2, this demonstrates the benefit of having the additional visual information for speech understanding.

Audio only and audio-visual speech intelligibility varied between talkers as expected. For clear speech in the audio only condition there was a difference of 15dB between the most intelligible and least intelligible of the four talkers and for the audio-visual condition a difference of 10dB. For vocoded speech, for Talkers 1, 2 and 3, performance also varied. The difference between the highest and lowest performing talkers being 6dB in the audio only condition and 5dB in the audio-visual condition. This demonstrates the significant variability in intelligibility between talkers in line with previous research (e.g. Cox et al., 1987; Gagne et al. 1994; Hazan & Markham, 2004; Lachs & Hernandez, 1998). In addition, if Talker 4 is considered in the vocoded speech analysis, the variation between the best and worst talkers extends to 17dB in audio only condition and 12dB in audio-visual condition. This is driven by the particular unintelligibility of Talker 4 in vocoded speech. The relative intelligibility of the target talker used in research would therefore potentially change the outcome of any project that was assessing audio or audio-visual performance, and the amount of visual speech benefit gained by participants. The same talker should be used throughout to ensure accurate measurement of performance levels is realised. The relative intelligibility of the target talker should also be assessed to ensure consistency between studies.

Variability in testing stimuli would also be particularly problematic for testing cochlear implant (CI) users. If it is assumed the vocoding of stimuli in this experiment is an accurate proximity of the type of audio information that CI users receive, variation in target talkers would significantly change hearing assessment results. It is especially concerning if outcomes from procedures or surgeries for CI users are assessed using varied stimuli; accurate testing of CI users should use standardised testing stimuli. If testing stimuli differ, comparison across time for patients and between patient groups becomes problematic. Given the large variation in patients' outcomes (Pisoni et al., 2016) and the importance of CI implants for patients' long-term health outcomes (Vermeire et al., 2005) it is important to ensure accurate and consistent measurement of performance. This is important for managing CI users' expectations and to assess the relative benefits of the implants.

Lachs and Hernandez (1998) and Gagne et al. (1994) found variability between talkers in different modalities. In Experiment 3, the target talkers' audio only and audio-visual performance levels were related in clear speech. Better audio only performance resulted in better audio-visual performance for each talker. However, there was some variability when audio only speech and visual only conditions are compared. Although the visual only condition was carried out with a different participant group to the audio only condition, the visual only results were better for Talker 2 than Talker 3, but Talker 3 had better audio only and audio-visual performance levels than Talker 2. Moreover, in the vocoded speech conditions, Talker 1 had the best audio only performance level but Talker 3 the best audio-visual performance level. Therefore, in line with Lachs and Hernandez (1998) and Gagne et al. (1994) talkers performance level vary somewhat as a result of the presentation modality. This is therefore important to consider if assessment of performance is being carried out in different presentation modes. There may be some variation in performance levels between audio and visual conditions because of differences in the stimuli which are therefore not attributable to the participant being tested.

Green et al. (2007) and Bent et al. (2009) argue that talkers rank similarly in terms of intelligibility in clear and vocoded speech. This is in line with the results of Experiment 3. Audio only performance levels were in the same rank order in the clear speech condition as the vocoded speech condition. This suggests audio intelligibility levels are independent of speech degradation; intelligibility levels across talkers do not vary with speech type. This is useful for research and assessment using a range of normal hearing and hearing-impaired populations. Audio intelligibility of the target talker used for one population sample would be equivalent for a different population sample.

There was also more variation in performance levels in vocoded speech conditions than in clear speech conditions, for both audio only and audio-visual tasks across all talkers. This is shown by the width of the whiskers and error bars (compare Figures 5.2 and 5.3). This is in line with Experiment 2 and previous research (Rosen et al., 2013) where more variation was found in vocoded speech conditions than in clear speech conditions. Talker 4 was the most unintelligible talker in clear and vocoded speech and had the greatest variation in performance levels. The width of the error bars for Talker 4 in clear speech ranges over 16dB (Figure 5.3). Talker 4 was removed from the main analysis for vocoded speech because of failed adaptive tracks for seven of the participants. The analysis for the remaining 17 participants in vocoded speech showed an audio only average performance level of 18.41dB with a standard deviation of 7.03. This large standard deviation shows the large variation in performance in vocoded speech for Talker 4.

5.5.2 Visual Speech Benefit

The amount of visual speech benefit received by participants was in line with previous research (e.g. Middelweerd & Plomp, 1987, Stacey et al. 2016). The amount of benefit ranged from 2.4dB to 7.4dB in clear speech for all four talkers, and between 3.6dB and 5.5dB in vocoded speech for Talkers 1, 2 and 3.

As hypothesised, overall, participants received more visual speech benefit in vocoded speech than in clear speech. This is in line with the predictions made if the PofIE were to apply, and also in line with other research; more benefit was found in degraded speech conditions than in clear speech conditions (Stacey et al. 2016). However, increased levels of visual speech benefit in vocoded speech were not found for all of the target talkers. There was no significant difference found between the amount of visual speech benefit in clear and vocoded speech for Talker 2. There was also no difference found for the target talker used in Experiment 2. This suggests the PofIE does not apply to all target talkers when comparing clear and vocoded speech. For some target talkers, degradation of the speech by vocoding leads to improved multisensory integration, however, other talkers do not provide the same beneficial increase. This is also similar to the results of Experiments 1 and 2 and other previous research (e.g. Tye-Murray et al., 2010), where the PofIE did not apply monotonically when the background noise used in the research varied. Therefore, overall, this would suggest the PofIE does not apply as monotonically to speech perception as proposed by Sumby and Pollack (1954) but has a more complex relationship with the integration of auditory and visual information. Variation in the stimuli, target talker or background noise, changes the nature of the speech integration.

For clear speech and vocoded speech, the amount of visual speech benefit generally increased as the intelligibility of the target talkers decreased. This is in line with the PofIE. As the audio information stream was degraded by the reduced intelligibility of the talker, the multisensory benefit from the audio-visual information increased. However, this relationship did not hold true for all talkers in vocoded speech. Overall, the most intelligible talker in vocoded speech was Talker 3 (mean audio only level of -1.23dB) and a mean visual speech benefit of 5dB. Talker 1 had a mean audio level of 1.5dB but had an average visual speech benefit of 3.6dB. Talker 3 provided greater visual speech benefit than Talker 1 despite being the more intelligible talker. Similarly, the talker used in Experiment 2 had an audio only performance level of 7dB in vocoded speech and an average visual speech benefit of 4dB. This suggests that for vocoded speech, the amount of visual speech benefit gained is not monotonically related to intelligibility levels for all target talkers. The monotonic relationship appears to be limited to more intelligible talkers and in less degraded speech. The monotonic relationship was observed in clear speech for all target talkers but not for vocoded speech. This may indicate there is a limit to the benefit that the PofIE confers. This in line with previous research (Altieri and Wenger, 2013; Ross et al. 2006), where maximal limits to the amount of visual speech benefit that could be gained were found.

There was also greater variation in the amount of visual speech benefit received in vocoded speech than in clear speech as shown by the width of the error bars (Figure 5.3). This suggests, as found in Experiment 2, performance in vocoded speech conditions is more varied than in clear speech conditions. As argued in section 4.9.1, greater variation in vocoded speech conditions may reflect participants' relative inexperience with vocoded stimuli. Vocoded stimuli may represent a greater challenge for participants and so the audio and visual information streams may be integrated in a different way to clear stimuli.

To address if integration varies as a result of inexperience with the stimuli, research could assess if experienced cochlear implantees (CI) gain different levels of visual speech benefit than naive normal hearing participants presented with vocoded speech (assuming vocoded speech stimuli simulates the type of speech experienced by CI users). This would establish if similar levels of benefit are found in degraded speech conditions regardless of the level of experience with degraded speech. This would be in line the proposed model of speech perception outlined in the PofIE; greater benefit from multisensory input is gained where unisensory input is degraded. Processed speech may provide equal levels of visual speech benefit regardless of the hearing characteristics of the population sample; normal hearing presented with vocoded speech and CI users would gain similar levels of benefit regardless of the level of experience with the processed speech.

Alternatively, if a difference was found between the amount of benefit between normal hearing participants using vocoded speech and participants with a CI, it would suggest experience with the type of speech used in the stimuli changes the amount of visual speech benefit received. Less experience with vocoded stimuli for normal hearing listeners would result in greater visual speech benefit. Levels of benefit would therefore be determined by experience and not by the amount of degradation in the speech. This was examined by Rouger et al. (2007). Compared to normal hearing listeners, Rouger et al. (2007) found that CI users integrated audio and visual information more effectively and gained greater visual speech benefit. This suggests that degraded stimuli did not provide greater levels of visual speech benefit regardless of hearing status. This is in contradiction to the PofIE, whereby greater visual speech benefit would be expected in more degraded stimuli compared to clear speech stimuli. However, it is also possible that changes to the cortical networks of CI users also contribute to the way in which multisensory information is integrated and change the amount of visual speech benefit that is received. Therefore, comparison across hearing impaired and normal hearing groups may be problematic.

It has also been argued that the normal hearing listeners in Rouger at al. (2007) integrated the audio and visual information suboptimally compared to other normal hearing participant groups (Stacey et al. 2016). To examine this further, research could be carried out with normal hearing participants who had considerable experience of vocoded speech stimuli and comparisons made with participants with no previous experience of vocoded speech. This would assess if increased visual speech benefit is found in degraded speech regardless of experience levels, in line with the PofIE.

In addition, as the results of Experiment 3 highlight, different talkers have different levels of intelligibility and so generate different levels of visual speech benefit. Comparison of target talkers could also be carried out to address whether integration varies by experience and by target talker. It is possible that participants, with repeated exposure to a particular target talker, may gain improvements in the amount of visual speech benefit for that talker. Research has found that listeners gain cross modal benefits from prior exposure to target talkers (Rosenblum et al., 2007, Sanchez et al., 2013). It is therefore feasible that with prior exposure to target talkers, differences in the levels of visual speech benefit would be found. This may vary as a result of the intelligibility of the target talker. Talkers who have particularly low levels of intelligibility may provide a greater increase in performance levels as participants become more familiar with the individual characteristics of that talker. Experience with vocoded stimuli for normal hearing listeners is examined in Chapter 7, where participants are given additional exposure to vocoded speech stimuli.

Changes to the amount of visual speech benefit gained as a result of changes to the target talker could also be examined if the background noise type varies. The results of Experiment 2 showed there was no variation in the amount of visual speech benefit gained if the numbers of background talkers changed for one target talker. Research could assess if the same results would be found if the target talker varied. It could be argued that the talker used in Experiment 2 was relatively unintelligible. Chapter 8 examines this and compares the intelligibility of the talkers used in Experiment 3 to the talker used in Experiment 1 and 2. It could be argued that if a different target talker was used with varying background noise types, the amount of visual speech benefit may differ.

Overall, the results of Experiment 3 show that as expected, the intelligibility of talkers varies significantly, and this changes the amount of visual speech benefit participants are able to gain from the stimuli. For clear speech, the amount of visual speech benefit increases as the audio and audio-visual intelligibility of the talker decreases. The results also highlight considerable variation in performance levels between participants, particularly in vocoded speech. Variation between participants is considered in the next two chapters. The amount of visual speech benefit gained is considered in relation to individual differences between participants that may explain some of the variation in performance levels found in Experiments 1, 2 and 3.

5.6 Chapter Summary

- Talkers are known to vary in audio intelligibility, but it is not known whether this variability changed the amount of visual speech benefit received by participants
- In Experiment 3, four target talkers were assessed for audio only, audio-visual, visual only performance levels and the amount of visual speech benefit they provided

- Significant variation in intelligibility between talkers was found in audio only and audio-visual conditions
- In clear and vocoded speech, the amount of visual speech benefit increased as the audio intelligibility of the target talker decreased in clear speech, in line with the Principle of Inverse Effectiveness
- There was a greater amount of visual speech benefit provided in vocoded speech than in clear speech overall, but this did not hold for all target talkers
- This suggests the Principle of Inverse Effectiveness, whereby multi-sensory integration is enhanced as unimodal performance declines, may not apply monotonically to visual speech perception tasks and may vary depending on the test stimuli used in the research.

Chapter 6 - Individual Differences Between Participants

Chapter 6 examines Experiment 4, which assesses individual differences between participants in the amount of visual speech benefit gained. Experiments 1, 2 and 3 showed a large variability in performance levels between participants but the reasons for this large variability are generally unknown. Experiment 4 examines differences in performance levels for both clear and vocoded speech. Individual difference measures assessed are; sensitivity to temporal fine structure information, aspects of attention, working memory capacity, ability to detect audio and visual synchrony, tendency towards general autistic traits and general speech performance levels. These individual difference measures are examined with respect to the variation in the amount of visual speech benefit received. The significance of these measures is discussed for both clear and vocoded speech.

6.1. Introduction

This chapter examines individual differences between participants that contribute to the amount of benefit gained from visual speech information. Experiments 1, 2 and 3 examined the changes to the amount of visual speech benefit received if the stimuli used in the research changed. These experiments also demonstrated the variability in the amount of visual speech benefit received by individual participants. For example, in Experiment 2, for the same target talker, the amount of visual speech benefit received in clear speech in a sixteen talker background noise ranged from a deficit of 3.1dB to a positive benefit of 11.6dB. Similarly, for vocoded speech in the same sixteen talker background noise, the range was from a deficit of 2.6dB to a positive benefit of 11.2dB. There is currently a lack of research explaining why this individual variability is so great. However, research has examined differences in visual only (lip reading), audio only, and audio-visual speech perception, and found some individual differences that contribute to task performance.

Research has found large individual differences in lip reading ability. Summerfield (1987), found performance levels varied between less than 10% correct to over 70% correct in tests of lip reading. Summerfield goes on to argue that lip reading skills are independent of some intelligence and verbal reasoning measures (e.g. Macleod & Summerfield, 1990) and are not improved significantly with training (e.g. Dodd, Plant & Gregory, 1989). Differences may be caused by disparities in the speed of low level neural processing (e.g. Shepherd, DeLavergne, Frueh, & Clobridge, 1977), but Summerfield (1987) concludes that it is not known why there are such large differences in lip reading abilities.

Other research has examined differences in audio only speech perception. Carbonell (2017) found that audio speech perception abilities for normal hearing listeners did not vary when the speech was degraded in differing ways, across a variety of speech tasks. They tested participants using three types of degraded speech; noise-vocoded, time-compressed, and speech in babble noise, and found participants performance levels correlated across the differing speech types. This suggests performance in degraded speech tasks is stable and therefore, other factors are influencing performance levels.

Heinrich, Henshaw & Ferguson (2015) found that for listeners with mild hearing loss, there was a range of -5dB to 11.25dB for the speech reception threshold when listening to speech in background noise. Speech performance was correlated with aspects of cognition (attention, memory, and non-verbal intelligence quotient) once hearing loss was partialled out, demonstrating that individual differences in listeners were affecting task performance.

It is not known whether the variability in visual only, audio only and audio-visual speech tasks are predicted by the same individual differences that may explain the variability in the amount of visual speech benefit received. There is currently a lack of research examining individual differences in the amount of visual speech benefit gained. Therefore, Experiment 4 examined individual differences in the amount of visual speech benefit received by participants. The individual difference measures chosen to try and predict the amount of benefit gained were based on previous research examining individual differences in visual only, audio only and audio-visual research. These measures are discussed in detail in section 6.2.

6.2 Experiment 4 - Individual Differences

6.2.1 General Speech Perception Ability

It is not known whether greater gains from visual information are found where participants are poorer or better at the audio only and audio-visual speech tasks. It could be argued that participants who are on average poorer at speech perception tasks may receive larger benefits from the visual speech information as they have greater scope for improvement in the task and therefore greater capacity to benefit from the addition of the visual information. Similarly, participants who are better at the speech tasks may have less room for improvement and so receive less benefit from the addition of the visual information. Conversely, it could be argued participants who are better at speech perception tasks have increased ability overall in understanding speech and so may gain more from visual speech information.

Therefore, to account for these possibilities, general speech performance ability was measured for each participant. General speech perception ability was measured by averaging the audio only and audio-visual performance for each participant for clear speech and for vocoded speech. This allowed for the initial level of expertise in the task for each participant to be considered when measuring the amount of visual speech benefit gained.

6.2.2 Temporal Fine Structure Information

Temporal fine structure information (TFS) refers to the rapid oscillations of soundwaves close to the centre frequency of the band (Lorenzi, Gilbert, Carn, Garnier & Moore, 2006; Moon & Hong, 2014). A detailed explanation of TFS can be found in section 2.1.3.

Sensitivity to TFS information is argued to play an important role in the ability to understand speech, and therefore sensitivity to TFS may predict the amount of visual speech benefit gained. For example, Strelcyk and Dau (2009) found that TFS performance levels highly correlated with audio only speech perception performance for closed set sentences in a two talker background noise, composed of one male and one female talker ($r= 0.84$, $p= .009$).

TFS information has also been shown to be especially important in background noise (e.g. Moore, 2008), where TFS information provides a cue to allow ‘dip listening’ i.e. listening more effectively in the gaps in fluctuating background speech noise (Lorenzi, Gilbert, Carn, Garnier, & Moore, 2006; Moore, 2008). It could therefore be argued that sensitivity to TFS information would be especially important in Experiment 4 as the speech perception tasks include background noise where TFS information is of greater importance.

Lorenzi et al. (2006) also argue that the loss of TFS information in their moderately hearing-impaired participants reduced their ability to benefit from ‘dip listening’ for speech in noise. Moore (2008) argues cochlea hearing loss reduces the ability to use TFS cues successfully. The vocoded speech type used in Experiment 4, and throughout this project, simulates the type of speech information experienced by CI users and therefore removes much of the TFS information. Therefore, the ability to use TFS information becomes especially important is the vocoded speech tasks in Experiment 4.

A test to measure sensitivity to TFS was included within Experiment 4 (TFS-LF 500Hz) (Hopkins & Moore, 2010). It was predicted that greater sensitivity to TFS information would lead to higher levels of visual speech benefit.

6.2.3 Attention

Attention is defined as the cognitive mechanism that selects sensory input for further processing (Talsma, Senkowski, Soto-Faraco & Woldorff, 2010). Selection can occur by top down means, based on internal intentions or relevance, or by bottom up means, whereby attention is directed by externally driven sources, often without voluntary control (Theeuwes, 1991). Being able to effectively select which source of information to process has been shown to be important in understanding speech, and therefore may predict the levels of audio-visual speech benefit gained.

Gibney et al. (2017) investigated the role of attention in multisensory integration using the McGurk task (McGurk & Macdonald, 1976). They reported that a higher perceptual load task (detecting the presence of a digit in an array of visual distractors) which diverted attention away from the task, resulted in less McGurk illusions. They argue this provides evidence that attention is necessary in order to successfully integrate audio and visual speech stimuli. Gibney et al. (2017) also examined the role of individual differences in the McGurk illusion and found differences across participants. For example, participants who did not perceive the McGurk illusion in the task without cognitive load, did not perceive the illusion in the low load or high load conditions. Therefore, for some participants, levels of speech integration were not affected by the attention task. This suggests that not all individuals are affected in the same way by attentional demands and therefore attentional differences may affect speech perception abilities and so levels of visual speech benefit gained.

It has also been argued that selective auditory attention is necessary to separate the relevant speech sounds in ambiguous auditory scenes, such as amongst competing talkers (Bizley & Cohen, 2013; Schoof & Rosen, 2015). Dai, Best and Shinn-Cunningham (2018) argue that auditory attention operates as a gain mechanism by enhancing the target stimuli and suppressing other extraneous distractions. Therefore, when attempting to identify speech within background noise, it could be argued that attention has an important role to play.

Heinrich, Henshaw and Ferguson (2015) measured another aspect of attention, focused attention, using the telephone search task (subtest 6) from the Test of Everyday Attention (TEA) (Robertson, Ward, Ridgeway, & Nimmo-Smith, 1994). They found a significant correlation between performance on the focused attention task and auditory only sentence perception in 8Hz modulated noise. Similarly, Schoof and Rosen (2015) measured auditory speech perception performance of older adults and individual differences, including ability to switch attention, using the visual elevator task from the TEA (Robertson, Ward, Ridgeway, & Nimmo-Smith, 1994). They argue attention switching is important for understanding speech in noise, whereby attention needs to be directed from the distracting talkers to the target talker. They found no relationship between performance on the attention task and age. Schoof and Rosen (2015) go on to argue that the participants in the study had normal hearing and therefore, were less likely to have cognitive impairments (including attentional deficits) as typically shown by older adults with hearing impairments (Lin et al. 2013). This suggests that not all aspects of attention are important for audio speech understanding.

The research evidence highlighted suggests some attentional abilities may be important for perceiving speech, particularly in background noise, and this may extend to the amount of visual speech benefit gained. Therefore, two parts of the TEA (Robertson, Ward, Ridgeway, & Nimmo-Smith, 1994) were included in Experiment 4. The TEA (Robertson, Ward, Ridgeway, & Nimmo-Smith, 1994) consists of eight sub-sets devised to measured various aspects of attention. Two aspects of attention were measured using two tasks from the TEA; firstly, elevator counting with reversal, and secondly, the visual map search task (version A). The elevator counting with reversals task is designed to measure aspects of selective attention and sustained attention using audio stimulus and was therefore deemed most appropriate for use in Experiment 4. The second part of the TEA used was the visual map search task (version A). This task measures selective attention using visual stimuli and therefore deemed most suitable to measure visual aspects of attention. Both auditory and visual aspects were therefore measured, as ability in these areas is important in completing auditory and audio-visual speech perception tasks. It was hypothesised that better performance on the attention tasks would result in greater levels of visual speech benefit.

6.2.4 Working Memory Capacity (WMC)

Research using audio only stimuli and visual only stimuli has highlighted some associations between verbal working memory capacity and performance on speech perception tasks. However, results are inconsistent. There is very also little research assessing the relationship between WMC and the level of visual speech benefit gained.

The relationship between WMC and hearing impairment had been widely studied. For example, Rudner, Rönnberg and Lunner (2011) tested older adults with hearing aids on their ability to understand speech in background noise, consisting of steady state or modulated background noise. They found that working memory capacity, measured by a reading span test, was highly predictive of performance in the speech perception test for the hearing-impaired sample. Moossavi, Etemadi, Javanbakht, Bakhshi and Sharafi (2016) tested children with cochlear implants (CI) on their ability to understand speech in speech type background noise. They found a significant positive correlation (r=0.80) between the backward digit span test and single word speech perception in background noise. Moossavi et al. (2016) argue this points to the importance of WMC in the ability to understand speech in noise for children with a CI. Similarly, Harris et al. (2013) found a more moderate but significant correlation (r=0.40) between speech and language test results in children with a CI and the Wechsler children’s backward digit span test (Wechsler, 1991).

Research assessing the relationship between WMC and understanding speech in noise has also been conducted with normal hearing participants. McCreery, Spratford, Kirby and Brennan (2017) assessed speech recognition performance for children in a steady state masking noise. They used single words, high predictability sentences and low predictability sentences as the target stimuli. Higher WMC capacity predicted better speech in noise recognition across all three speech types; single words, high predictability sentences and low predictability sentences.

Millman and Mattys (2016) assessed middle-age and older adults with normal hearing (mean age 53.5) on speech perception in modulated background noise. They tested speech perception performance using IEEE sentences, as used in this research project, using five different female target talkers. No assessment of the intelligibility of the target talkers is mentioned (see Chapter 5 for analysis of the variation in intelligibility of talkers). Millman and Mattys (2016) found no relationship between performance on the backward digit span test and speech perception. However, they argue the backward digit span test not only measures the phonological aspect of working memory but also requires executive control; the digits require processing in order that they can be repeated back to the experimenter in reverse order. Therefore, they argue the backward digit span test is not predictive of speech in noise performance. However, as discussed in Chapter 5, these results maybe confounded by the difference in intelligibility between the target talkers. Similarly, Füllgrabe and Rosen (2016) conducted a meta-analysis assessing the relationship between WMC, measured by the reading-span test, and speech in noise performance using a sentence test with modulated or unmodulated background noise, for adults ranging in age (18–91 years). They found WMC declined with age (r=0.59) as did performance on the speech in noise task (r=0.68). However, further analysis showed that the correlation between WMC and audio speech perception was only significant for the adults over 40 years of age. They argue, therefore, that WMC is not predictive of speech in noise performance in younger normal hearing participants.

Rosemann et al. (2017) tested the WMC of normal hearing adults, when listening to vocoded speech sentences. They tested WMC capacity using the Automated Operation Span Test (Unsworth, Heitz, Schrock & Engle, 2005). In this test, participants were asked to solve mathematical problems while retaining letter information. They established a significant correlation between WMC and understanding of vocoded speech in normal hearing listeners (r= 0.47).

Research has also looked at the relationship between WMC and lip-reading performance in normal hearing participants. Working memory capacity has been argued to be a significant predictor of lip-reading performance (e.g. Lidestam, Lyxell & Andersson, 1999; Lyxell & Holmberg, 2000; Picou, Ricketts & Hornsby, 2011). However, again, these results vary. A younger adult sample (mean age 20.6) and an older adult sample (mean age 76.8) were assessed on their WMC using simple and complex letter span tasks, and their spatial working memory using a grid recall task, in research by Feld and Sommers (2009). A significant proportion of the variance in lip-reading ability was explained by performance on the spatial working memory task for both the younger and older adults. Spatial working memory explained 36% of the variance in performance in lip reading ability overall. This suggests similar methods are being used by both age groups to achieve better lip-reading performance. However, no relationship between the WMC task and lip-reading performance was discovered, although there were significant correlations between performance on the WMC tasks and the spatial working memory tasks.

Overall, the relationship between WMC and speech perception performance is inconsistent and confounded by differences in test stimuli, in both the measures used to assess WMC and in the speech perception task. For the purposes of Experiment 4, it was deemed most appropriate to measure WMC using the backward digit span test (Weshler, 2008). The backward digit span test is part of the Wechsler Adult Intelligence Scale (WAIS) (Weshsler, 2008) which measures general IQ. The backward digit span test measures working memory capacity (WMC) and auditory processing, which is most similar to the speech perception tasks in Experiment 4; hearing a sentence for three seconds and repeating the recalled words at the end of that sentence. Therefore, a relationship between levels of verbal WMC measured by the backward digit span test (Wechsler, 2008) and performance on speech perception tasks could be expected. It was predicted that higher working memory capacity would lead to greater levels of visual speech benefit gained.

6.2.5 Audio-Visual synchrony (AVS)

Audio-visual synchrony (AVS) refers to whether two events, an auditory and a visual input, are simultaneous or successive. Research has highlighted differences in ability to recognise synchrony. Therefore, a synchrony task was included in

Experiment 4 to ascertain if there was any relationship in the ability to identify synchrony and the level of visual speech benefit gained.

Despite differences in the temporal transmission of visual and auditory stimuli (i.e. light travels quicker than sound through air, visual stimuli is processed more slowly than auditory stimuli), the human brain is adept at synchronising these two events into a coherent whole and thereby understanding speech (Vroomen & Keetels, 2010). Research has demonstrated that participants who were better lip-readers were more able to detect asynchrony (McGrath & Summerfield, 1985). Conversely, Conrey and Pisoni (2006) found no relationship between lip reading performance and ability to detect synchrony. Conrey and Pisoni (2006) did, however, find a positive correlation between AV speech perception performance and auditory only performance and ability to detect AV synchrony. Conrey and Pisoni (2006) used City University of New York (CUNY) Sentences Test (Boothroyd, Hanin & Hnath-Chisolm, 1985), for their speech perception task, which are similar to the IEEE sentences used within this research project, and time-reversed degraded speech. Baskent and Bazo (2011) established a similar positive relationship between AV and AO stimuli and ability to detect AV synchrony for normal hearing participants in steady speech-shaped background noise. They used closed set meaningless sentences as the target stimuli. This body of research demonstrates the relationship between AV synchrony and speech perception performance for understanding sentences, using degraded speech and within background noise.

Research has also highlighted differences in the ability to identify temporal difference, or asynchrony, between audio and visual stimuli. For example, Summerfield (1987) argues professional musicians have a greater ability to detect synchrony between audio and visual presentation of music, when the delay is as short as 30ms. The typical level at which asynchronous stimuli is thought to be noticeable is argued to be 40ms (McGrath & Summerfield, 1985), although the width of the window across which participants believe stimuli to be synchronous varies across individuals (Baskent & Bazo, 2011; Donohue, Darling & Mitroff, 2012; Stone et al., 2001). Vatakis and Spence (2006) found an increased ability to detect AV synchrony in speech over music using a student population sample who were non-musicians.

Taken together, this research suggests expertise may influence the ability to detect AV synchrony, supporting the importance of individual differences in completing this task. It can also be argued therefore, that participants who are better at the speech tasks generally may be considered to be ‘experts’ at speech perception and therefore have increased ability to detect AV synchrony in speech. This is supported by Conrey and Pisoni (2006) who found that participants who were better at identifying synchrony in auditory leading speech did not display improved accuracy in detecting synchrony in the non-speech conditions, suggesting that exposure and experience in dealing with speech leads to expertise status. It is argued by Conrey and Pisoni (2006) that this reflects the typical experience of speech, whereby visual motion often precedes auditory input, and therefore stimuli where the auditory input precedes the visual input is unnatural and more easily identified. It has long been established that visual leading asynchronies are more difficult to identify than auditory leading asynchronies, (e.g.; Conrey & Pisoni, 2003, 2006; Dixon & Spitz, 1980; Grant, Wassenhove & Poeppel, 2003), again reflecting the typical speech experience of the visual motion preceding the auditory input.

Overall, this evidence suggests that sensitivity to AV synchrony may predict speech perception performance and the levels of visual speech benefit gained. Therefore, in order to assess the relationship between sensitivity to AV synchrony and visual speech benefit, an AV synchrony task was included in Experiment 4. The task was to identify if single words spoken by a single female talker were synchronous or asynchronous. Stimuli were presented with the visual leading the auditory stimulus by 185ms (V185A) in order that sensitivity could be established using a more difficult synchrony test. Conrey and Pisoni (2006) reported approximately 0.75 proportion synchronous responses at an asynchrony level of approximately V185A. It was hypothesised that better ability in the AV synchrony task would lead to higher levels of visual speech benefit.

6.2.6. Autism Spectrum Quotient (AQ)

The tendency towards greater autistic traits was measured using the Autistic spectrum quotient (AQ) (Baron-Cohen, Wheelwright, Skinner, Martin & Clubley, 2001). The AQ was designed as a self-reporting descriptive tool, demonstrating an individual’s tendency towards greater autistic traits as defined by the score on the

AQ. Therefore, the AQ operates as a continuum, across those with a clinical diagnosis of autism and the general population (Lai, Lombardo, Chakrabarti & Baron-Cohen, 2013; Mayer 2017).

Autism is a highly variable developmental disorder. Those with a diagnosis range in function from high to low with a wide range of symptoms across the spectrum. The three main behavioural symptoms are; difficulty with social interactions, communication difficulties and repetitive behaviours (Miles, 2011). It is thought to be caused by a combination of genetic and environmental factors (Chaste & Leboyer, 2012; Miles, 2011). Research has highlighted specific speech perception deficits shown by those with a diagnosis of autism, that may extend to those who score more highly on the AQ.

Stevenson et al. (2016) argue that the integration or ‘binding’ of multisensory information, known to deliver gains, such as gains in audio-visual speech perception (Sumby & Pollack, 1954), is suboptimal in those with a diagnosis of an autism spectrum disorder (ASD). They found a wider temporal binding window (TBW) for ASD individuals which was specific to complex speech stimuli (McGurk stimuli), in contrast to more simple stimuli; flashes and beeps, dynamic handheld tools. Stevenson et al. (2017) assessed children between the ages of 7 and 16 on their speech perception abilities (single nouns, single female talker, multi-talker babble background noise). They found that ASD individuals were significantly poorer in the speech perception tasks than the typically developing children in the audio-visual condition and the lip reading condition. However, there was no significant difference in performance in the auditory only condition. This suggests deficits are apparent in the visual domain or in the integration of the audio and visual information. Smaller TBW’s in typical populations have been shown to lead to increased susceptibility to the McGurk illusion (Stevenson, Zemtsov & Wallace, 2012). This suggests there may be a relationship between the size of the TBW and integration of audio visual-speech.

Evidence has also been found of infants as young as 10 months old, who were later diagnosed with autism, displaying atypical responses to AV synchrony (Falck-Ytter, Nyström, Gredebäck, Gliga & Bölte, 2018). Falck-Ytter et al. (2018) used eye tracking to assess preferential looking towards targets where the audio and upright animation were in synchrony, in comparison to stimuli where the animation was inverted and played in reverse. Those infants who were later diagnosed with autism did not distinguish between the two types of stimuli, suggesting that these infants process multisensory information sub optimally when compared to typically developing infants.

Research has also examined speech perception in background noise for individuals with an ASD diagnosis. An adolescent population was tested in the research of Smith and Bennetto (2007). They measured the number of key words correctly identified in sentences in a four talker background noise type, using an adaptive track procedure, similar to the procedure used in this research project. They found the typical population were better lipreaders than the ASD sample, had significantly lower SRT's in the audio-visual conditions, and interestingly, gained a greater benefit from the addition of the visual information in audio-visual speech perception. Similarly, Alcántara, Weisblatt, Moore and Bolton (2004) compared task performance in identifying words in sentences for typical and ASD participant samples. They used varying types of background noise; a single female talker, steady speech-shaped noise; speech-shaped noise with temporal dips, steady speech-shaped noise with regularly spaced spectral dips and speech-shaped noise with temporal and spectral dips. They found significant differences between groups in the conditions where there were temporal dips in the background noise. The ASD group were less able to benefit from the temporal dips, when the target speech can be more easily 'glimpsed', resulting in poorer speech perception in noise for the ASD a participant sample.

Taken together, this suggests that individuals with autism may be less sensitive to asynchronous stimuli and have deficits in audio-visual speech perception. Research by Donohue et al. (2012) suggests this deficit in multisensory processing could be extended to those without a diagnosis of autism but who exhibited higher autistic traits as measured by the AQ. Donohue et al. (2012) found a significant correlation ($R = -0.30$, $p = 0.002$) between judgements of simultaneity, using tonal auditory stimuli and a checkerboard visual image, and score on the AQ. Their participants were healthy adults without a diagnosis of autism. A higher score on the AQ lead to a bias towards auditory first presentation as being seen as more synchronous. They argue that these results may indicate deficits in temporal processing shown by participants with higher self-reported autistic traits or alternatively, may illustrate impaired multisensory processing. In order to examine this further, Donohue et al. (2012) conducted a second experiment to establish if judgments of synchrony using visual-visual stimuli would yield a significant correlation with scores on the AQ. No significant correlation was found. They suggest this points to deficits in multisensory processing in individuals with higher scores on the AQ rather than temporal processing deficits. The positive correlation between scores on the AQ and the simultaneity judgment task also remained when other individual difference measures known to contribute to multisensory perception abilities (video game expertise & ADHD) were taken into consideration. However, as Donohue et al. (2012) point out, an audio-audio task was not conducted in order to rule out temporal processing deficits in the auditory domain.

Research has also found a relationship between autistic traits in the general population, measured by the AQ, and the ability to discriminate between voices. Skuk, Palermo, Broemer, and Schweinberger, (2017) found a negative correlation between voice recognition of familiar voices and autistic traits in an adolescent population sample. Skuk et al. (2017) found that those with higher autistic traits were less able to discriminate familiar voices from unfamiliar voices, highlighting discrimination deficits in auditory processing. This may result in greater difficulty identifying the target speech of a familiar voice among other completing talkers for those who have higher autistic traits; the voices would sound more similar leading to greater discrimination difficulty. This is in contrast to a typical listener who would be able to identify the target voice more easily. By extension, this may also apply to the audio and audio-visual speech perception tasks included in Experiment 4. Those with higher autistic traits may not benefit from building up familiarity with the target voice to aid speech perception, as has been shown to benefit typical listeners (Nygaard & Pisoni,1998) (see Chapter 5 on the effects of changing the target talker). Interestingly, in Experiment 1 in Nygaard and Pisoni (1998), listeners achieved between 28% and 97% correct identification of the talker, after nine days of training and building familiarisation with the talkers. This may be explained by individual differences in the ability to discriminate between the voices predicted by levels of autistic traits as outlined in Skuk et al. (2017). The levels of autistic traits in participants was not measured by Nygaard and Pisoni (1998) in their study.

Overall, this substantial body of research highlights some of the challenges and deficits to speech perception, particularly in background noise, faced by those with a diagnosis of autism. What is less clear, however, is whether this deficit extend to individuals who report higher levels of autistic traits as measured by the AQ. As outlined, Donohue et al. (2012) argue there may be deficits in multisensory processing in individuals with higher scores on the AQ. However, there is very little research that has examined speech perception performance or the benefits of visual speech information in relation to autistic traits in the general population. Therefore, Experiment 4 included the AQ in order to examine the relationship between scores on the AQ and the level of benefit received from visual speech information. It was predicted that participants who scored more highly on the AQ would receive less benefit from the visual speech information.

6.2.7 Hypotheses for Experiment 4

It was hypothesised, in line with previous experiments in this project, that clear speech would be easier to understand than vocoded speech, and that audio-visual performance would be better than audio only performance. It was also hypothesised, that as found in Experiment 3, there would be more visual-speech benefit in vocoded speech than in clear speech for this target talker.

The number of measures chosen was large and therefore this experiment was intended to be exploratory because of the constraints of obtaining a suitably large enough participant sample size within the current project.

The target talker chosen for target stimuli in Experiment 4 was Talker 3 from Experiment 3. Talker 3 was chosen because this talker showed the largest variability in overall performance in Experiment 3 and was therefore considered to more suitable to demonstrate variability in performance in Experiment 4.

For each of the individual difference measured outlined in section 6.2, the following effects were hypothesised for clear and vocoded speech in terms of the regression model;

- General speech perception ability; it was not known whether greater general speech perception ability would predict greater or lesser levels of visual speech benefit.

- Temporal fine structure information; greater sensitivity to TFS information would lead to higher levels of visual speech benefit.
- Attention; better performance on the attention tasks would lead to higher levels of benefit from the visual speech information.
- Working memory capacity; greater working memory capacity would lead to higher levels of visual speech benefit.
- Audio-visual synchrony; increased ability to detect AV synchrony would lead to higher levels of visual speech benefit.
- Autism spectrum quotient; greater tendency towards autistic traits, indicated by a higher score on the AQ, would lead to reduced levels of visual speech benefit.

6.3 Method

6.3.1 Design

A within-participants design was used in Experiment 4. audio only and audio-visual speech reception thresholds (SRT_{50}) were measured in decibels (dB) for clear and vocoded speech.

The first analysis was a 2 (modality: audio only vs audio-visual) x 2 (speech type: clear vs vocoded) ANOVA. The dependent variable was SRTs measured in dB.

The second analysis was a regression analysis. The outcome variable was visual speech benefit (AV-AO SRTs, in dB for clear or vocoded speech). The predictor variables were general speech perception ability in dB averaged across the AV and AO tasks, score on the TFS test, score on the attention tasks, verbal working memory capacity, score on the AV synchrony task and score on the AQ.

6.3.2 Participants

Twenty-five participants were recruited from the student population at Nottingham Trent University. The increased sample size needed to test the large number of variables in Experiment 4 was not possible to achieve within the constraints of the project. Therefore, it was intended that the project be exploratory in nature, allowing that a larger sample size would be needed to explore the results further. Therefore, the number of participants recruited was determined by the criteria outlined in section 4.3.2.

The mean age of the participants was 23, age range 19 to 34, with six being male. All other criteria were the same as Experiment 1. Participants were recruited from the student population at Nottingham Trent University and were rewarded with research credits. Consent was obtained from each participant as agreed by the Nottingham Trent University Research Ethics Committee. Participants also confirmed normal hearing, normal or corrected to normal vision and had English as their first language.

6.3.3 Materials and Procedure

6.3.3.1 Equipment

The experiments were conducted in a multi person IAC Acoustics 40a-5 audiology booth situated at Nottingham Trent University. The stimuli were presented in an identical manner to Experiment 2.

6.3.3.2 Target Stimuli

The sentences were chosen from the IEEE corpus (Rothauser et al., 1969). Eighty sentences were presented, spoken by a single male talker, Talker 3 from Experiment 3. Talker 3 was chosen because the results of Experiment 3 showed this talker had the largest distribution of results in dB across the audio and audio-visual conditions. This allowed an increased probability of finding variance in the results in the speech tasks in Experiment 4. Twenty IEEE sentences were used in the practice task, as outlined in section 4.3.3, five practice sentences for each condition.

6.3.3.3 Background Stimuli

Background noise for use in the experiment was created using the UCL speaker database (Markham & Hazan, 2002). Sixteen talkers (8 male and 8 female) were used as the background noise. This was the same background noise file used in Experiment 3.

6.3.3.4 Speech Processing

The vocoded speech was processed in an identical way to the previous experiments in this project (see section 4.3.3.4 for a full explanation of the procedure used).

6.3.3.5 Predictor Variables

General Speech Perception Ability

The general procedure for the practice task and the main experiment for this experiment were identical to that presented in Experiment 2. Following the practice task, each participant was presented with the four conditions in a random order; audio only for clear and vocoded speech, audio-visual for clear and vocoded speech. Each condition was assigned a different list of twenty sentences from the 80 IEEE sentences recorded and each list was presented in a random order. Therefore, each participant had different sentences for each condition and in a different presentation order.

General speech perception ability was measured using the target stimuli as outlined in 6.3.3.2. Audio only and Audio-visual Speech Reception Thresholds(SRT_{50}), measured in decibels (dB) were obtained for clear and vocoded speech. The average of these were calculated for clear speech and for vocoded speech. The average was the general speech perception performance for each participant for clear speech in dB and for vocoded speech in dB.

Temporal Fine Structure Perception (TFS)

Sensitivity to temporal fine structure (TFS) information at low frequencies was measured using the binaural TFS test (TFS-LF 500Hz) (Hopkins & Moore, 2010). The visual stimuli were presented on a Dell PC monitor measuring 26cm x 41cm and the auditory stimuli presented over HD280pro headphones (Sennheiser, Wedemark, Germany) via a custom built digital-to-analogue converter. The test involves a two-alternative forced choice task. Participants are asked to identify, by clicking on A or B using the mouse, which of two groups of four tones contained the interaural phase difference (IPD), which sounds as if the tones are moving within the head (laterization shift). In one of the two groups of tones, the second and fourth tones had an IPD. The tones were played at a frequency of 500Hz and at 30dB sensation level. Each tone was 400ms long and separated from the next tone by 20ms, the gap between the two sets of tones was 500ms. The initial IPD was presented at the maximum 180° and varied using a two-up one-down adaptive procedure until six reversals were achieved. The TFS threshold was calculated by measuring the geometric mean of the six reversals, which corresponded to the 71% correct point.

Attention

Two aspects of attention were measured from the Test of Everyday attention (TEA) (Robertson et al., 1994); elevator counting with reversal, visual map search task (version A). The elevator (lift) counting with reversals involved an audio recording of a series of tones, a middle-pitched tone indicates the lift is moving. If a high tone is heard, this indicates the lift is changing direction and is going up and if a low tone is heard, this indicates the lift is changing direction again, and is going down. The participant listens to a series of tones, establishing internally whether the elevator has gone up or down following each tone. These were played over HD280pro headphones (Sennheiser, Wedemark, Germany) via a custom built digital-to-analogue converter, as used in the speech perception tasks. At the end of each series of tones, the audio playback requests that the participant verbally reports on which floor they have calculated the lift has stopped. This is recorded by the experimenter. There were ten trials which get more difficult by the inclusion of more tones in each series. Scores were recorded as the number of correctly identified floors and scored out of a maximum of ten.

The second task was the visual map search (version A) (Robertson et al., 1994). The task involved searching for restaurant symbols (a knife and fork) on a map and circling the symbols when found. The task was timed over two minutes. The number of found restaurant symbols in two minutes was totalled and scored out of a possible maximum 80 correct.

Verbal Working Memory Capacity

Verbal working memory capacity was measured using the backward digit span test (Wechsler, 2008). The experimenter read out loud a series of digits, one per second, and participants were asked to verbally repeat back the digits in reverse order. The task started with three digit numbers and the number of digits increased to a maximum of 8 digits, with two trials in each. The task was stopped when the participant repeated two consecutive sequences of digits incorrectly. The task was scored as numbers of sequences correctly repeated, out of a maximum score of 16.

AV Synchrony (AVS)

Ability to detect audio-visual synchrony was measured using seventy words from the English Lexicon Project (Balota et al., 2007), spoken by a single female talker (Figure 6.2). The words used were two syllable words of mid-range frequency. Example words being, bottle, jumper, filter, sugar, happy.

Figure 6.2 An example of the stimuli used in the audio-visual synchrony task.

Sentences were presented using E-Prime software (Version 2.0 Psychology Software Tools Inc., Sharpsburg, US). A practice block of ten words was presented in a random order before the main task; five words in synchrony and five words out of synchrony. In the main task, sixty words were presented in a random order; 30 words in synchrony and 30 words that were asynchronous. Stimuli were presented with the visual leading the auditory stimulus by 185ms (V185A) in order that sensitivity to identifying AV synchrony could be established using a more difficult synchrony test (Conrey & Pisoni, 2006).

Participants were seated at approximately 50cm from the screen and asked to indicate whether a single word was in synchrony or out of synchrony by pressing M or Z on the keyboard. The visual stimuli were presented at a size of 24cm high x 41cm wide on a Dell PC monitor measuring 26cm x 41cm. The audio stimuli were presented in an identical manner to the audio files in the speech perception task. Scores were calculated out of a possible 60 correct words and converted to the percentage correct.

Autism Spectrum Quotient (AQ)

Tendency towards greater autistic traits was measured using the Autism spectrum quotient (AQ) (Baron-Cohen et al., 2001). The questionnaire consists of 50 questions and are answered on a four-point Likert scale; definitely agree, slightly agree, slightly disagree and definitely disagree. Examples of the questions include, ‘I find social situations easy’, and ‘I notice patterns in things all the time’. The 50 questions include ten questions assessing each of five different areas; social skill, attention switching, attention to detail, communication and imagination. These five areas are argued to be indicative of autistic-like behaviour (Baron-Cohen et al., 2001). A higher score on the AQ indicates a greater tendency towards autistic traits; a score of more than 32 indicating individuals who have clinically significant levels of autistic traits. Ruzich et al. (2015) conducted a systematic review of published research and established that the mean AQ score across the general population was 16.94 (95% CI 11.6, 20.0) and for those with an autism diagnosis was 35.19 (95% CI 27.6, 41.1), validating the original suggested scoring criteria. However, it is worth noting that Baron-Cohen et al. (2001) found that those with a scientific background, particularly mathematicians, score significantly higher than those with a humanities or social sciences background.

Participants were instructed to read each question carefully and circle how strongly they agree or disagree with each statement by circling the answer. Scores were totalled, giving an AQ score for each participant out of possible maximum of 50. A higher score on the AQ indicates greater self-reported levels of autistic traits.

6.4 Results

6.4.1 Audio and Audio-Visual Performance

SRT_{50s} were calculated in the same way as Experiment 2, by averaging the last ten trials in each participant’s adaptive track. All adaptive tracks were successful, with performance being between 30% and 70% correct as expected. Figure 6.3 shows the SRT’s obtained in each of the AO and AV tasks for clear and vocoded speech.

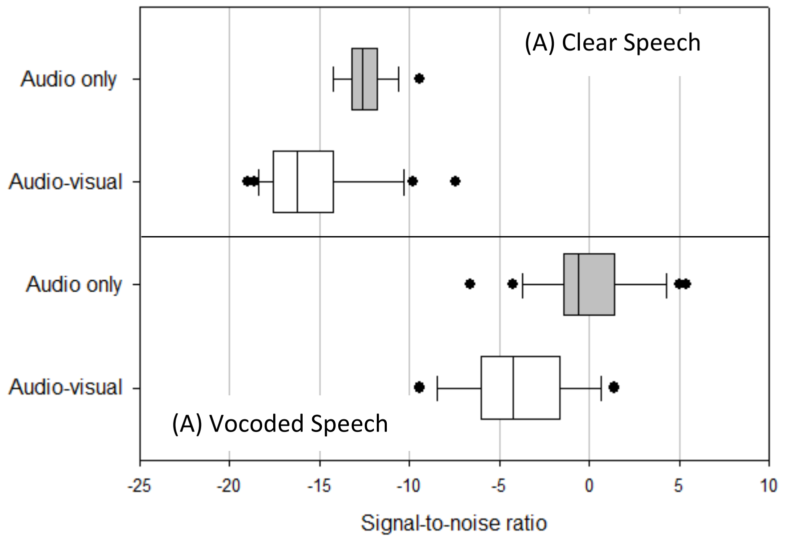


Figure 6.3. Audio only and audio-visual Speech Reception Thresholds for Clear (Panel A) and Vocoded (Panel B). The whiskers show the 10-90% range and the black dots show outlier data which fall outside that range.

A 2 (modality: audio only vs audio-visual) x 2 (speech type: clear vs vocoded) ANOVA showed a significant main effect of modality ($F(1,24) = 38.29$, $p<0.001$, $\eta_p^2 = 0.62$). Performance was better in audio-visual than audio only conditions. A significant main effect of speech type ($F(1,24) = 619.37$, $p<0.001$, $\eta_p^2 = 0.96$) was also found. Clear speech was easier to understand than vocoded speech. There was no interaction between modality and speech type ($F(1,24) = .42$, $p=.52$, $\eta_p^2 = .017$).

6.4.2 Visual Speech Benefit

The benefit received from the visual speech information was calculated by measuring the difference in between AV and AO conditions. For clear speech, the mean visual speech benefit was 3.29dB (SD=3.01) and for vocoded speech was 3.71dB (SD=3.53). There was no significant difference between the amount of visual speech benefit received between clear and vocoded speech ($t(24)=.52$, $p=.61$).

The variation in the amount of visual speech benefit received across participants was noticeably greater in the vocoded speech task than in the clear speech task. Figure 6.4 shows the distribution in the amount of visual speech benefit received for both clear and vocoded speech.

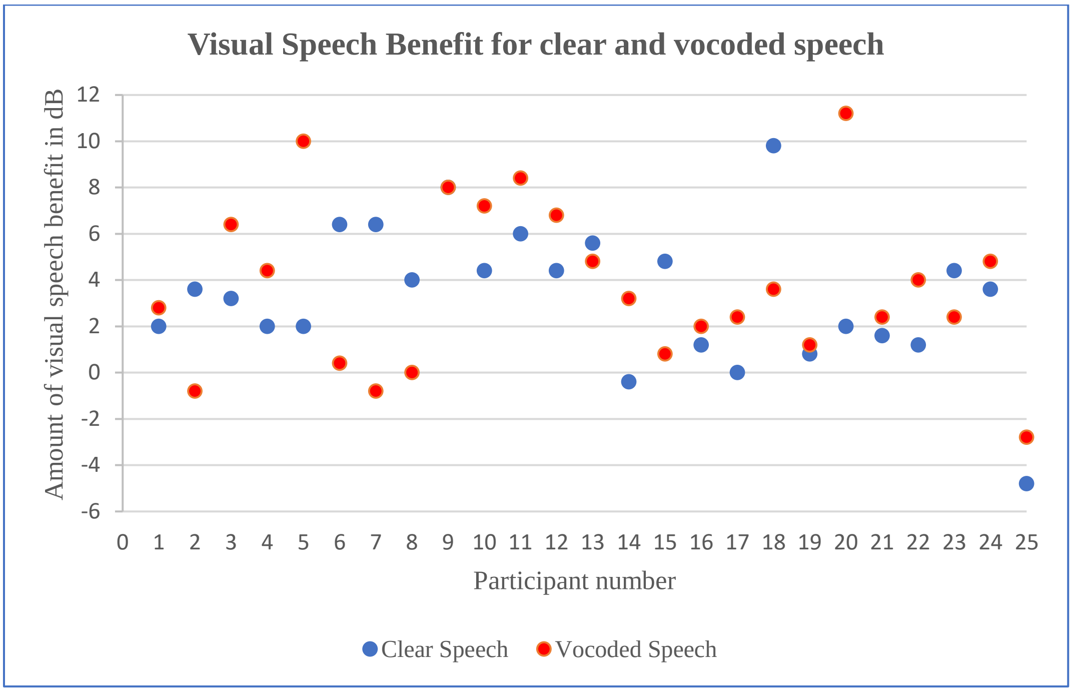


Figure 6.4 The distribution of the amount of visual speech benefit received for clear and vocoded speech.

6.4.3 Individual Difference Measures

Table 6.1 shows the means and standard deviation for the individual difference measures in Experiment 4.

As outlined in section 6.4.1, general overall performance was better in the clear speech conditions than the vocoded speech conditions ($t(25)=24.89$, $p<.001$). The mean performance in the TFS task (M=24.30, 500Hz, 30dB sensation level) was in line with expected performance levels on this task (Hopkins & Moore, 2010). Similarly, performance on the attention tasks were within expected levels (Robertson et al., 1994) as was the performance level on the working memory capacity test (Wechsler, 2008). The AV synchrony task was designed by the researcher so no direct comparison with other research can be made. The mean score on the AQ was 16.12, which compares favourably to a general population mean of 16.94 (Ruzich et al., 2015).

Table 6.1 Means and standard deviation for the individual difference measures in Experiment 4.

Individual difference measures		Mean	SD
General Speech perception	CLEAR	-13.90dB	1.71
	VOCODED	-2.02dB	2.20
Temporal fine structure (TFS)		24.30	12.50
Attention -Elevator with reversals (TEA-EL)		7.48	2.40
-Visual map search (TEA-MAP)		70.44	6.44
Working memory capacity (WMC)		8.64	1.98
Audio-visual synchrony (AVS)		77.07%	12.29
Autism spectrum quotient (AQ)		16.12	7.98

6.4.4 Regression Analysis

A regression analysis was run to ascertain if the levels of visual speech benefit received were predicted by the individual difference measures; average overall general speech perception ability, score on the TFS task, scores on the attention tasks, working memory capacity, score on the audio-visual synchrony task and score on the AQ. Some assumptions of the regression analysis were met; the data was normally distributed and had homogeneity of variance, tolerance and VIF were within acceptable levels for both the clear and vocoded speech. However, working memory capacity significantly correlated with other predictor variables. Table 6.2 shows the results of the correlational analysis. Working memory capacity (WMC) significantly correlated with the audio-visual synchrony task (AVS) (r=.46), with one of the attention tasks; elevator counting with reversals (TEA-EL) (r=.46), and with the general speech perception performance in the vocoded condition (VOC) (r=-.41). Working memory capacity was therefore removed from the main analysis.

Table 6.2 Correlations between the individual difference measures

Variables	CLE	VOC	TFS	TEA-EL	TEA-MAP	WMC	AVS
VOC	0.28						
TFS	0.30	-0.05					
TEA-EL	0.16	-0.06	0.18				
TEA-MAP	0.12	0.01	-0.26	0.04			
WMC	-0.21	-.41*	-0.18	.46*	0.12		
AVS	-0.04	0.04	-0.07	0.31	0.19	.46*	
AQ	-0.04	-0.24	0.01	0.20	-0.02	0.14	0.23

*$p \leq .05$

Regression analysis for clear speech

The regression analysis was carried out to ascertain if the individual difference measures, including overall performance on the speech perception tasks, predicted levels of visual speech benefit. However, it is worth noting that the results should be treated with some caution. Modelling of the data suggests that if there is unequal variance between the AO and AV performance levels then some correlation with visual speech benefit is to be expected, given that visual speech benefit is calculated by subtracting the audio visual performance level from the audio-only performance level.

For clear speech the model was significant. The results of the multiple regression analysis showed that for clear speech, three of the predictors explained 60% (adjusted R^2= .60) of the variability in the amount of visual speech benefit gained (F(6,18)=7.1, p=.001) (Table 6.3). Firstly, general speech perception performance was a significant predictor. Participants who had a better performance on the audio only and audio-visual tasks received more benefit from visual speech (β=-.59, p<.001). Secondly, participants who were more able to detect AV asynchrony obtained more visual speech benefit (β=.44, p=.016). Finally, participants who scored more highly on the AQ, indicating a greater tendency towards autistic traits, gained less benefit from visual speech information (β =-.36 p=-.016).

Table 6.3 Results of the regression analysis for clear speech

Variables	B	SE(B)	β	t	p
General speech perception	-1.05	.25	-.59	-4.26	<.001
TFS	-.03	.04	-.14	-.96	.35
TEA-EL	-.14	.18	-.11	-.78	.45
TEA-MAP	-.05	.07	-.11	-.79	.44
AVS	.11	.04	.44	3.09	.016
AQ	-.13	.05	-.36	-2.66	.016

Regression analysis for vocoded speech

For vocoded speech the model was not significant. There were no significant predictors of the amount of visual speech benefit received, with -.24% of the variance explained (adj. R^2 -.24%, F=.24, p=.96). Table 6.4 shows the results of the regression analysis for vocoded speech.

Table 6.4 Results of the regression analysis for vocoded speech

Variables	B	SE(B)	β	t	p
General speech perception	-.15	.38	-.92	-.39	.70
TFS	-.03	.07	-.11	-.44	.67
TEA-EL	-.11	.36	-.07	-.29	.78
TEA-MAP	-.01	.13	-.03	-.11	.92
AVS	-.02	.07	-.08	-.33	.75
AQ	-.08	.11	-.18	-.74	.47

Regression Analysis for Working Memory Capacity

A separate analysis was carried out to ascertain if there was a relationship between working memory capacity (WMC) and the amount of visual speech benefit gained. Working memory capacity was removed from the main analysis because of significant correlations with other predictor variables, the audio-visual synchrony task (AVS), with one of the attention tasks (elevator counting with reversals) and with the general speech perception performance in the vocoded speech condition. The regression analysis showed that WMC did not predict the amount of visual speech benefit gained for clear speech (adj. R^2 .08%, F=.19, p=.67), or for vocoded speech (adj. R^2 -.03%, F=.07, p=.79).

6.4.5 Performance over time on the speech perception tasks

As outlined in section 6.4.2, the distribution of the amount of visual speech benefit in the vocoded condition was greater than that in the clear speech condition. In addition, the regression model for vocoded speech was a very poor fit to the data (adj. R^2 -.24%, F=.24, p=.96). It is possible that listeners may not have gained enough experience with the vocoded stimuli in order to process the vocoded stimuli in the same way as the clear speech stimuli, leading to greater variation in the vocoded speech tasks.

In order to assess if participants were processing the vocoded stimuli in a different way to the clear speech stimuli in Experiment 4, change in performance over the course of 20 sentences was examined for clear speech and for vocoded speech. This was calculated by subtracting each participant's starting SNR from the SNR of each of the subsequent 20 sentences. This gave a value representing how far the SNR has moved from its starting value for each sentence. These values were averaged for each of the 20 sentences, giving a mean value per sentence. Figure 6.5 shows the change relative to the starting SNR for each participant for clear and vocoded speech.

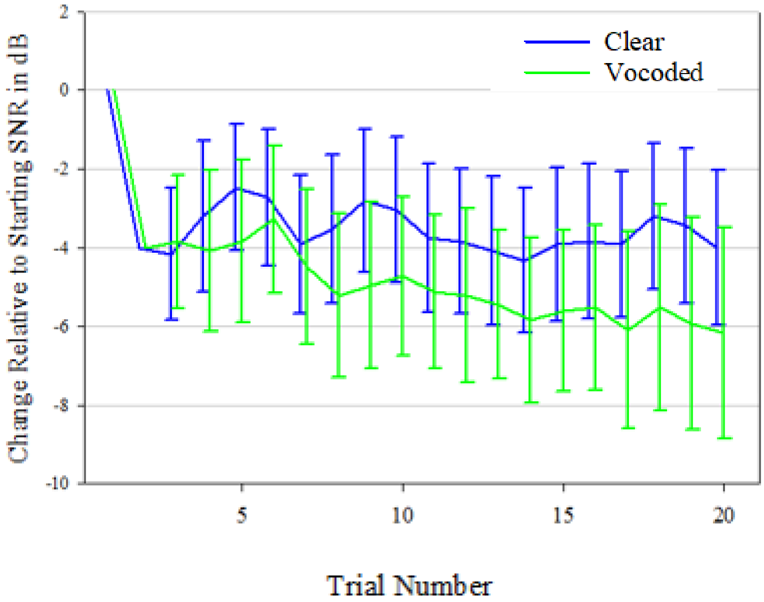


Figure 6.5 Change relative to the starting SNR for each participant over the 20 sentences in the clear speech and vocoded speech audio only conditions in Experiment 4. Error bars show 95% confidence intervals.

Figure 6.5 shows that over the course of the 20 sentences, the vocoded speech condition moved further away from the starting SNR than in clear speech. On average, participants moved 6.2dB away from their starting SNR in vocoded speech in contrast to 4dB in clear speech. In addition, the error bars in the vocoded speech condition are wider than for clear speech. This suggests there is more variability in performance over time in the vocoded speech condition.

6.5 Discussion

6.5.1 Overall Results

In line with previous experiments in this project, it was expected that in Experiment 4, clear speech would be easier to understand than vocoded speech and audio-visual speech would be easier to understand than audio only speech. The results of Experiment 4 were in line with this expectation, evidencing the difficulty of understanding vocoded speech in comparison to clear speech and the benefit of having the visual information to the auditory input when perceiving speech. In Experiment 4 the amount of visual speech benefit did not differ significantly in the clear and vocoded speech conditions as hypothesised, however, there was a larger variation in the amount of benefit received by participants in the vocoded speech conditions than the clear speech conditions. This is in line with the results of Experiment 1, 2 and 3, where more variation in the vocoded speech conditions than the clear speech conditions were found.

The primary aim of Experiment 4 was to ascertain if individual difference measures predicted the levels of visual speech benefit gained by participants. Overall, the regression model was significant for clear speech, with three of the predictor variables explaining 60% of the variance in the level of visual speech benefit; general speech perception ability, ability to detect audio-visual synchrony (AVS) and score on the autism spectrum quotient (AQ). The other predictor variables measured, sensitivity to temporal fine structure information, and attention, did not significantly predict levels of visual speech benefit gained as hypothesised. The model was not significant for vocoded speech with none of the predictors explaining the amount of visual speech benefit gained by participants. The result of the regression analysis in relation to the individual difference measures is discussed in detail in section 6.5.2. for clear speech and in section 6.5.3 for vocoded speech.

6.5.2 Individual Difference Measures - Clear Speech

6.5.2.1 General Speech Perception Ability

In Experiment 4, higher levels of visual speech benefit were predicted by better overall performance on the task in terms of audio only and audio-visual performance. The audio and audio-visual performance levels were averaged for each participant. Therefore, better audio only understanding of speech in background noise, in addition to being better able to use the visual information in the audio-visual condition, led to greater levels of visual speech benefit. Macleod and Summerfield (1987) argued that being able to benefit from visual information, in terms of lip reading, led to significantly greater levels of visual speech benefit. Experiment 4 developed this further. In addition to being better able to make use of the visual information in audio-visual speech tasks to gain higher levels of visual speech benefit, better audio only performance led to higher levels of visual speech benefit. This suggests participants who have speech in noise expertise as evidenced by a better average performance in the audio and audio-visual speech tasks, receive greater benefit from the visual speech information in the audio and audio-visual modalities; unisensory (audio only) and multisensory (audio-visual integration).

6.5.2.2 Audio-Visual Synchrony (AVS)

Ability to detect AV synchrony was also predictive of gains in visual speech benefit; greater sensitivity to AV synchrony predicted higher levels of visual speech benefit. This supports Conrey and Pisoni (2006) who found a positive correlation between AV speech perception performance and auditory only performance and ability to detect AV synchrony. Experiment 4 found the ability to detect AV synchrony predicted the amount of visual speech benefit gained. This suggests the same mechanisms that allow participants to detect AV synchrony are being used to gain benefits from visual speech information. However, Conrey and Pisoni (2006) used CUNY sentences (Boothroyd et al., 1985) in time-reversed degraded speech. CUNY sentences are easier to understand than the IEEE sentences used in Experiment 4 and have greater contextual cues allowing greater predictability (Kong, Donaldson & Somarowthu, 2015). An example of a CUNY sentence being, ‘I have a sore throat and a very bad cough’. They also used time-reversed degraded speech, which is more difficult to understand than clear speech, but did not use background noise. Baskent and Bazo (2011) found ability to detect AV synchrony predicted performance on audio only and audio-visual speech tasks, in steady speech-shaped background noise. They used closed set meaningless sentences as the target stimuli.

The results of Experiment 4, along with the research of Conrey and Pisoni (2006) and Baskent and Bazo (2011) suggest the ability to detect AV synchrony is important in varied speech types, using differing stimuli and in quiet and in varied types of background noise. This would appear to underline the importance of the ability to detect AV synchrony for speech perception generally; ability to detect AV synchrony leads to speech expertise. Summerfield (1987) found synchrony expertise for music over speech in a musician sample and Vatakis and Spence (2006) found an increased ability to detect synchrony in speech over music in a non-musician sample. This would suggest expertise is possible to acquire through experience, such as the extensive training received as a musician. Alternatively, it could be argued such expertise is innate, if individuals are drawn to be musicians because of an innate ability. Nevertheless, this points to the importance of individual differences in the ability to detect AV synchrony.

6.5.2.3 Autism Spectrum Quotient (AQ)

A lower score on the AQ predicted higher levels of visual-speech benefit in Experiment 4. Research has demonstrated those with an ASD diagnosis have specific speech perception deficits in the audio-visual and visual only domain (Stevenson et al., 2017). The results of Experiment 4 suggest that speech specific deficits in multi-sensory processing shown by those diagnosed with ASD, may be extend to non-clinical populations who score more highly on the AQ.

Individuals with a higher score on the AQ have been shown to have difficulty discriminating between voices. Skuk et al. (2017) found that those with higher autistic traits were less able to discriminate familiar voices from unfamiliar voices, highlighting discrimination deficits in auditory processing. In Experiment 4, it is possible those with a higher score on the AQ would not benefit from building familiarity with the target talker during the course of the experiment, and would therefore have had greater difficulty discriminating between the target talker and the background talkers.

Other research (Smith & Bennetto, 2007; Alcántara et al., 2004) found speech specific deficits in ASD individuals, Experiment 4 suggests this may extend to those with a higher score on the AQ. Smith and Bennetto (2007) found typically developing adolescents had lower SRT's than ASD individuals listening to speech in a four talker background noise and also gained more benefit from the visual speech information. Experiment 4 suggests the deficit in speech perception and benefit from visual speech information shown by individuals with an ASD, may extend to individuals who score more highly on the AQ but who do not have a diagnosis of ASD. Similarly, Alcántara et al. (2004) found significant deficits for ASD individuals when listening to speech with temporal dips in the stimuli. If this speech perception deficit extends to individuals with a higher score on the AQ, it is possible listeners in Experiment 4 with higher AQ scores received less benefit from the temporal dips in the background noise. However, the background noise used in Experiment 4 had 16 talkers and therefore there were very few temporal dips available to listeners.

6.5.2.4 Other Non-Predictive Variables

The other predictor variables measured, sensitivity to temporal fine structure information and attention, did not significantly predict levels of visual speech benefit gained as hypothesised. Additionally, the separate analysis for working memory capacity did not predict levels of visual speech benefit gained. These are discussed in the next section.

Temporal Fine Structure Information (TFS)

Sensitivity to temporal fine structure information (TFS) was hypothesised to lead to greater benefit from visual speech information. This was not found in Experiment 4. Strelcyk and Dau (2009) found that TFS performance highly correlated with audio only speech perception performance for closed set sentences in a two talker background noise. In addition, it has been argued TFS information is of particular importance in background noise to aid ‘dip listening’ (Lorenzi et al., 2006; Moore, 2008). However, in Experiment 4 the background noise consisted of 16 talkers in a continuous steam of information. Therefore, it could be expected there would be very few dips in the background noise that would benefit speech perception. Similarly, Strelcyk and Dau (2009) used a two talker background noise type, whereby there may have been dips in the audio steam of information. This would suggest therefore that in a background noise type of many talkers, as used in Experiment 4, ability to use TFS information more successfully is not as beneficial as in other noise types. The background noise may be too challenging for enhanced ability to use TFS information to prove useful. Further research could examine variation in the amount of benefit received from TFS information if the type of background noise changes.

Attention

In Experiment 4, two measures of attention were used from the Test of Everyday Attention (TEA) (Robertson et al., 1994); elevator counting with reversal, and the visual map search task (version A). The results of Experiment 4 showed that the tests of attention were not predictive of the amount of visual speech benefit gained.

Schoof and Rosen (2015) found no relationship between auditory speech perception and older adults and attention. Experiment 4 found no relationship between attention, measured by the elevator task with reversals (auditory selective and sustained attention), and the visual map search task (visual selective attention). This suggests aspects of attention measured in Experiment 4 are not predictive of visual speech benefit gained for younger adults (mean age 23, range 19-34) in addition to the older adult sample of Schoof and Rosen (2015).

However, attention has been shown to be important in multisensory integration, using the McGurk illusion (Gibney et al., 2017). It has also been argued that selective auditory attention is necessary to separate the relevant speech sounds in ambiguous auditory scenes, such as amongst competing talkers (Bizley & Cohen, 2013; Schoof & Rosen, 2015). It could be argued that the type of background noise used in Experiment 4, 16 talker noise, was significantly challenging that attention was captured by the stimuli. Smaller attentional differences, measured by the TEA tasks, were not captured because of the difficulty of the task in comparison to simpler tasks, such as the McGurk illusion (McGurk & Macdonald, 1976). Therefore, any differences in attention that may be present between participants were not great enough to contribute to speech perception performance and to the amount of visual speech benefit gained.

Working Memory Capacity (WMC)

It was predicted that higher WMC would lead to greater gains in visual speech benefit. However, WMC significantly correlated with the audio-visual synchrony task and with one of the attention tasks (elevator counting with reversals) and was therefore removed from the main analysis. The separate regression analysis for this variable, found WMC did not predict the amount of visual speech benefit gained for clear or vocoded speech.

Previous research has established a relationship between hearing impaired individuals and WMC. Rudner at al. (2011) found a significant relationship between WMC, measured by a reading span test, and speech perception in noise. Moossavi et al. (2016) and Harris et al. (2013) found a significant relationship between speech and language test results in children with a CI and the backward digit span test. However, in normal hearing subjects research is inconsistent. McCreery et al. (2016) found a significant relationship between WMC and speech in noise understanding for children. In contrast, Millman and Mattys (2016) found no relationship between speech perception using IEEE sentences and WMC measured suing the backward digit span test. Similarly, Füllgrabe and Rosen (2016) found no relationship between WMC and speech perception for adults under 40 years of age. The results of Experiment 4 support the research of Millman and Mattys (2016) and Füllgrabe and Rosen (2016) as WMC was not predictive of the amount of visual speech benefit gained. This body of research suggests WMC may be predictive of speech benefits for hearing impaired subjects and for children, but is not significant in a normal hearing younger adult population.

6.5.2.5 Overall Model for Clear Speech

It is worth noting the sample size used in this experiment is limited, so any conclusions drawn regarding analysis should be considered to be exploratory. However, taken together the existing data suggests the significant predictors of visual speech benefit suggest a type of visual-speech in noise expertise. Greater abilities on audio only and audio-visual speech perception, ability to detect AV synchrony and a lower score on the AQ make a significant contribution to being able to extract the most benefit from being able to see the face of the talker. This suggests a type of ‘super visual-speech perceiver’, whereby expertise in general speech performance and in detecting AV synchrony, and a lower score on the AQ, lead to better overall ability to obtain visual speech benefit.

It is not known whether this ‘super visual-speech perceiver’ hypothesis extends to other types of stimuli and different types of background noise. Experiments 1 and 2 highlighted the differences in visual speech benefit if the number of talkers in the background varied and Experiment 3 highlighted the differences if the target talker varied in intelligibility. Further research could examine the ‘super visual-speech perceiver’ hypothesis if the speech stimuli varies. It is not known if the expertise suggested by the results of Experiment 4 is as a result of the particularly difficult background noise type, 16 talker background noise, and if this expertise extends to other types of background noise. It could be argued that in less challenging background noise types there is more opportunity to benefit from glimpses in speech and, therefore, it becomes less necessary to rely on expertise in speech perception as the task is less challenging; expertise is only necessary in particularly challenging tasks. Future research could address this by repeating Experiment 4 with different types of background noise and with more or less intelligible talkers than used in Experiment 4. This could establish at what point the 'super visual-speech perceiver's' expertise becomes useful. This also may apply to the vocoded speech stimuli. This is discussed further in section 6.5.3.

It is also not known is if visual speech perception expertise is influenced by where participants are looking on the face of the talker or by differences in multisensory integration. Greater benefit from audio-visual stimuli and increased ability to detect AV synchrony may be influenced by the gaze behaviour of the participants. Equally, expertise on audio and audio-visual tasks and ability to detect AV synchrony may be affected by differences in multisensory integration between participants. The 'super visual-speech perceiver' hypothesis could be supported by either argument. Experiment 6 will examine this further by using eye tracking to assess gaze direction and duration for participants and if this is predictive of the amount of visual speech benefit gained. This is discussed in detail in Chapter 7.

6.5.3 Individual Difference Measures - Vocoded Speech

It was hypothesised that the predictor variables would significantly contribute to the levels of visual speech benefit gained for vocoded speech. There were no significant predictors of the amount of visual speech benefit gained for vocoded speech.

Overall performance in the vocoded speech tasks was significantly worse than the clear speech tasks and there was more variation in the vocoded speech tasks than the clear speech tasks. As shown in Figure 6.4, the amount of visual speech benefit was more widely distributed than the amount of benefit in clear speech. Additionally, Figure 6.5 shows the change in audio only performance relative to the starting SNR for clear and vocoded speech. This shows that performance in the vocoded speech condition changed more over the course of the 20 sentences than the clear speech condition. Performance improved for vocoded speech more than for clear speech. This suggests listeners are improving in performance because of more exposure to the vocoded stimuli. Moreover, the error bars for vocoded speech are wider than for clear speech. This demonstrates the increased variability in the vocoded speech condition than in the clear speech condition. In addition, it could be argued that greater variability in the vocoded speech tasks means the results are less predictable and more unstable. Results are less likely to be predicted by other variables if performance is unstable.

Research has highlighted that performance in vocoded speech tasks improves over time for normal hearing listeners (e.g. Davis, Johnsrude, Hervais-Adelman, Taylor & McGettigan, 2005; Rosen, Faulkner & Wilkinson, 1999). Therefore, in Experiment 4, listeners may not have gained enough experience with the vocoded stimuli in order to process the vocoded stimuli in the same way as the clear speech stimuli.

In Experiment 4, there was a practice before the main experiment as used in Experiments 1, 2 and 3, to afford familiarity with the vocoded stimuli. However, because of the reduced number of vocoded speech perception conditions in Experiment 4, greater familiarity with the vocoded stimuli may have been gained in Experiments 1, 2 and 3. Experiments 1 and 2 included six audio and audio-visual vocoded speech conditions and Experiment 3 included eight audio and audio-visual speech conditions. In Experiment 4 there were four vocoded speech conditions.

Therefore, overall it could be argued that greater experience with the vocoded stimuli would change the performance levels of participants. Without prior exposure and learning of the vocoded stimuli, performance is unstable and does not reflect multisensory integration of the stimuli as designed, but reflects lack of experience with the vocoded stimuli. Rosen et al. (2013) found similar results with clear speech stimuli and vocoded background noise, whereby participants had little experience with the stimuli before the experiment and would therefore benefit from further training to improve task performance.

Consequently, it could be argued that as the vocoded speech task was more challenging than the clear speech task, speech expertise could not be utilised. As discussed in section 6.5.2.5, a ‘super visual-speech perceiver’ model was suggested for clear speech. It may be that the ‘super visual-speech perceiver’ hypothesis only applies to clear speech tasks, where years of experience with clear speech have provided substantial amounts of learning and experience. Alternatively, it may be possible that innate abilities contribute to visual speech performance and therefore, with training, expertise with vocoded speech may be acquired and the ‘super visual-speech perceiver’ be realised in the vocoded speech model. This is discussed in more detail in Chapter 7, where more experience with the vocoded stimuli is included within Experiment 5.

6.6 Chapter Summary

- There is significant variation in the amount of visual speech benefit received between participants
- Individual differences that may predict the variation in the amount of visual speech benefit gained were measured in Experiment 4. The individual difference measures were; general speech perception ability, sensitivity to temporal fine structure information, aspects of attention, working memory capacity, ability to detect audio and visual synchrony, and tendency towards general autistic traits
- For clear speech, three of the predictors explained 60% of the variance in the amount of visual speech benefit received. The significant predictors were; general speech perception performance, ability to detect audio and visual synchrony and tendency towards autistic spectrum traits
- For vocoded speech, there were no significant predictors of the amount of visual speech benefit gained
- There was no difference in the amount of visual speech benefit gained between clear and vocoded speech
- Performance in vocoded speech tasks was more varied than in clear speech tasks, in line with the previous experiments in this project
- There was greater change in performance over time in the vocoded speech tasks than the clear speech tasks
- The greater variation in vocoded speech tasks and greater change over time in the vocoded tasks suggests less experience with vocoded speech may be contributing to variation in performance levels. Greater experience with clear speech may be leading to more stable performance levels.

Chapter 7 – Individual Differences and Eye Tracking

Chapter 7 examines the final experiment in this project; Experiment 5. The significant predictors of the amount of visual speech benefit found in Experiment 4; general speech perception performance, ability to detect audio-visual synchrony and tendency towards autistic traits are re-examined with the addition of eye tracking data to assess gaze direction and duration. Gaze behaviour is added as a predictor to the model to ascertain if gaze behaviour also predicts the amount of visual speech benefit participants gain. This is examined for clear and vocoded speech. A familiarisation task is also included for the vocoded speech conditions in order that greater stability in performance levels in vocoded speech can be found.

7.1. Introduction

Chapter 7 re-examines the contribution of individual differences to the amount of visual speech benefit received by listeners. Experiments 1, 2 and 3 demonstrated the variability in the amount of visual speech benefit received by individual participants. Experiment 4 examined individual difference that may account for the variation in the amount of visual speech benefit gained. The results of Experiment 4 showed that 60% of the variation in levels of visual speech benefit in clear speech were significantly predicted by three measures; ability on audio only and audio-visual speech perception tasks, ability to detect audio-visual (AV) synchrony and a lower score on the Autism spectrum quotient (AQ). Experiment 5, discussed in this chapter, tested listeners using the same speech perception tasks and significant predictors of the amount of visual speech benefit gained from Experiment 4. In addition, Experiment 5 measured eye gaze direction and duration in the audio-visual speech tasks using eye tracking.

It was suggested in Chapter 6 that some participants have an expertise in perceiving visual speech in noise; a ‘super visual-speech perceiver’. Better ability in general speech performance, in detecting AV synchrony, and a lower score on the AQ, lead to better overall ability to obtain visual speech benefit in clear speech. However, for vocoded speech, there were no significant predictors of the amount of visual speech benefit gained. It was highlighted that performance in the vocoded speech tasks was more variable, was less predictable and showed more improvement over the course of the task than the clear speech condition. Therefore, the results of the vocoded tasks may reflect lack of experience with the vocoded stimuli. Subsequently, participants would be processing the vocoded stimuli in a different way to the clear speech stimuli. In order to investigate this further, Experiment 5 included a familiarisation exercise to improve stability in the vocoded speech task. This discussed in section 7.2.

It was also argued that differences in gaze behaviour across participants may be contributing to variation in the amount of visual speech benefit gained. Expertise on audio-visual tasks and ability to detect AV synchrony may be affected by differences in multisensory integration between participants. Alternatively, where participants are looking on the face of the talker may affect gains in visual speech benefit. Gaze behavioural differences are also particularly important for those with a diagnosis of ASD, which may extend to those with a higher score on the AQ. Experiment 6 will examine this by using eye tracking to assess gaze direction and duration in the audio-visual speech tasks, and if gaze behaviour is predictive of the amount of visual speech benefit gained. This is discussed in detail in section 7.3.

7.2 Experiment 5

7.2.1 Individual Differences and Familiarisation with Vocoded Speech

Experiment 4 demonstrated there was more variability in the vocoded speech tasks than the clear speech tasks. Figure 6.5 shows the audio only change relative to the starting SNR for each participant and shows that the error bars in the vocoded speech condition are wider than in the clear speech condition. Furthermore, participants showed greater improvement over time in the vocoded speech than the clear speech; the ending SNR for participants was further away from the starting SNR in vocoded speech than in clear speech. This is also shown in Figure 6.5. Additionally, the distribution in the amount of visual speech benefit was more widely dispersed in the vocoded conditions than in the clear speech conditions. (Figure 6.4). This suggests listeners are improving in performance because of more exposure to the vocoded stimuli. Therefore, listeners may be performing differently in the vocoded speech task than the clear speech task because of lack of experience with the vocoded stimuli.

A number of research studies have assessed learning over time for listeners using vocoded speech (e.g. Hervais-Adelman, Davis, Johnsrude, Taylor & Carlyon, 2011;

Shannon, Zeng, Kamath, Wygonski & Ekelid, 1995; Stacey & Summerfield, 2007). For example, Rosen, Faulkner & Wilkinson (1999) tested normal hearing listeners on four channel shifted spectral speech, simulating the type of speech produced by a four channel cochlear implant (CI). Listeners were given a total of three hours experience with the audio-visual and then audio only stimuli. This resulted in performance improving for identifying key words in sentences from 1% at initial testing to nearer 40% at the conclusion of the training sessions. However, it is worth noting Rosen et al. (1999) tested four participants in this study. As shown in Experiment 4 in this project, performance on vocoded tasks shows considerable variation between participants, therefore, learning for Rosen et al.’s (1999) four participants may not be reflective of learning for all listeners.

Davis, Johnsrude, Hervais-Adelman, Taylor and McGettigan (2005) tested normal hearing listeners using six channel noise-vocoded speech over a series of experiments. Davis et al.’s (2005) first experiment demonstrated a significant improvement in identifying key words over the course of 30 sentences. This supports the results of Experiment 4 in this project as shown in Figure 6.5; performance improved over the course of the 20 sentences. However, Davis et al. (2005) argue that improvements shown in their first experiment may be as a result of gaining familiarity with the task procedure and not with perceptual changes in vocoded speech understanding. This is addressed in their second experiment. Davis et al. (2005) presented listeners with the vocoded sentence, the sentence in clear speech and finally, the vocoded sentence again. Results showed a significant improvement in identification of the vocoded sentence after the clear speech presentation was heard in comparison to a group of participants who did not have access to the sentence in clear speech. This suggests perceptual gains in vocoded speech understanding and not improvements in task familiarity were being made by listeners. In Davis et al.’s (2005) third experiment, the audio presentation of the sentence in clear speech was replaced by a written presentation of the sentence. Performance in this task was not significantly different to their second experiment; presentation of the sentence in clear speech or in written form had the same effect on sentence identification. Davis et al. (2005) argue this suggests improvements in learning noise vocoded speech are not only supported by low-level acoustic information but also by access to higher-level linguistic content of the sentences.

Davis at al. (2005) argue that exposure to a vocoded speech sentence a second time, after gaining knowledge of the sentence in clear speech or by seeing the written sentence, results in a dramatic ‘pop out’ effect and significant improvement in intelligibility on the second presentation of the vocoded speech sentence. This is particularly noted for speech in noise, as used in Experiment 5 in this project. Jacoby, Allan, Collins and Larwill (1988) presented sentences to listeners in white noise. Participants judged the background noise to be significantly louder when new sentences were heard, than when an old sentence was heard; the sentences were heard once before during an earlier part of the experiment. This demonstrates that familiarity with stimuli can have an immediate perceptual effect on listeners experience of the task.

Experiment 4 had a reduced number of vocoded speech conditions than in Experiments 1, 2 and 3 in this project, therefore, greater familiarity with the vocoded stimuli may have been gained in previous experiments. Consequently, it could be argued that greater experience with the vocoded stimuli would change the performance levels of participants; the vocoded speech task was more challenging than the clear speech task and so speech expertise could not be utilised. As discussed in section 6.5.2.5, a ‘super visual-speech perceiver’ model was suggested for clear speech. The ‘super visual-speech perceiver’ hypothesis may only apply to clear speech tasks, where substantive experience with clear speech may have provided significant levels of learning. Alternatively, it may be possible that innate abilities contribute to visual speech performance and therefore, with training, expertise with vocoded speech may be gained and the ‘super visual-speech perceiver’ be realised in the vocoded speech model.

To that end, a familiarisation task was included in Experiment 5 to improve listeners’ understanding of vocoded speech. This was designed so that participants would perform in a similar way in the vocoded speech tasks and the clear speech tasks in Experiment 5. The familiarisation task included 30 sentences. Significant improvement in understanding of vocoded speech has been shown over the course of 30 sentences (Davis et al., 2005). The familiarisation task involved presenting the sentence in vocoded speech, the sentence being written on the PC monitor and finally, the vocoded sentence being repeated (Davis et al., 2005). This is outlined in detail in section 7.3.3.5.

7.2.2 Individual Differences and Eye Tracking

7.2.1.1 General Speech Perception and Gaze Behaviour

When perceiving speech, there are three salient areas of the face that are the main focus of gaze; the mouth, the nose and the eyes (Buchan, Paré & Munhall, 2008). Gaze on these areas shifts as task demands change. For example, Buchan et al. (2008) found gaze was concentrated more on the eye area when the task involved emotional judgments but shifted to the mouth region when the task involved word identification only. This suggests gaze direction is modified to access the most relevant facial region for the task; the eye region contains important social information (Baron-Cohen, Wheelwright, Hill, Raste & Plumb, 2001) and the mouth region aids speech perception (Buchan et al. 2008).

Differential gaze patterns have also been observed for speech perception tasks that include background noise. Where speech perception becomes more difficult with the addition of background noise, focus shifts to the mouth area for longer periods to aid understanding (e.g. Buchan et al. 2008; Yi, Wong and Eizenman, 2013). Research suggests this is an adaptive procedure. MacDonald, Marchman, Fernald and Frank (2018) presented clear speech without background noise, and in background noise consisting of brown noise at a SNR of 2.87dB. They were presented to adults and children (age 3-5). The speech task involved a real-time comprehension exercise. They found both adults and children spent more time focusing on the talking face in the noisy conditions than the clear speech condition, allowing more information to be gathered from the visual image. MacDonald et al. (2018) argue this shows similar adaptation to task demands in younger and adult listeners; increased background noise prompts increased focus on the source of the target speech, even at a very early stage of development.

Similarly, using clear speech, Vatikiotis-Bateson, Eigsti, Yano and Munhall (1998) assessed gaze patterns for audio-visual speech in background noise types of no noise, low, medium and high noise levels. The speech was monologues spoken by a single talker lasting 35-45 seconds. The background noise consisted of multi-lingual talkers and music recorded at a party. Two main areas of focus were established on the taking face; the mouth and eyes. The proportion of the trials gaze was directed to the mouth area increased with the background noise type, around 35% with no background noise to 55% when background noise was high. His demonstrates a shift in gaze fixations in speech in noise. Where speech perception is more challenging with increasing background noise, greater fixation is directed towards the mouth area to aid understanding.

However, this has been shown to vary as task demands change. Lusk and Mitchel (2016) also found gaze shifts towards the mouth area in difficult to understand speech. They used artificial language in clear speech as the stimuli. Participants looked at the mouth area for the majority of the time whilst trying to understand the language, but then moved their gaze away from this area as the artificial language became more familiar. This suggests the mouth area becomes a focus to increase perception but then attention may move elsewhere once familiarity with the task is realised.

Research has also examined levels of visual speech benefit gained by differential gaze patterns. Rennig, Wegner-Clemens and Beauchamp (2018) presented clear syllables without background noise, and sentences in -16dB pink noise, and tracked eye gaze differences between participants. They found that participants who spent more time focusing on the mouth region in the clear syllables condition received significantly more visual speech benefit, although the amount of benefit gained had considerable variation (ranging between 6% and 56%, mean 31%). This highlights the variation in gaze patterns between participants and the benefit of focusing on the mouth area to improve speech perception using this particular stimuli. In contrast to the clear syllables condition, the sentences in pink noise condition showed no relationship between gaze and the amount of visual speech benefit gained. This may reflect task difficulty in the speech in noise task; 93% of the time on average was spent looking at the mouth area with low variability between participants. This suggests the high SNR of -16dB used in this task was removing any individual variability in gaze behaviour; high task difficulty was producing near ceiling effects. It could be argued that if different SNR’s were used, gaze behaviour may have shown more variation and illustrated differences in the amount of visual speech benefit gained.

Alsius, Wayne, Paré and Munhall (2016) also found considerable variation between participants in the benefit received from visual speech information. They tested participants using words and sentences in a multi-talker babble background noise with and SNR's of -2 dB for words and -3 dB for sentences. The results showed that participants who gained the most visual speech benefit (identified as a high visual gain subgroup) spent more time looking at the mouth area than the participants who gained less visual speech benefit. The lower SNR's used by Alsius et al. (2016) of -2dB and -3dB may have afforded the differences in gaze patterns than those observed by Rennig et al. (2018) who used an SNR of -16dB.

Research has also looked at the relationship between eye movements and degraded speech, although the current quantity of research in this area is limited. Wagner, Toffanin and Baskent (2016) assessed gaze behaviour of normal listening participants in clear and degraded speech. The audio stimuli consisted of eight channel sine-wave vocoded speech, as used in this project. The task involved listening to the vocoded sentence and then clicking on the relevant picture on the monitor screen that was mentioned in the sentence. They found participants were slower to look at the correct picture in the vocoded speech condition, suggesting slower processing of the signal. Participants were also slower in giving a response in vocoded speech than in clear speech, suggesting increased difficulty in identifying the key word. This suggests that more effort would be required by participants to process vocoded speech and therefore may look at the mouth area for longer in order to aid speech perception. However, as Experiment 5 contains challenging background noise (16 talker), if may be expected that participants would look towards the mouth area in clear speech as well as in vocoded speech. Therefore, difference in the duration of focus on the mouth area may not be found; there may be near ceiling effects for both speech types.

Overall, this body of evidence suggests that preferential gaze towards the mouth area of a talking face aids speech understanding. In addition, participants who display this gaze behaviour gained more visual speech benefit. However, due to the challenging nature of the task it may be expected that there would be little difference in gaze behaviour between the clear and vocoded speech conditions.

7.2.1.2 Autistic Traits and Gaze Behaviour

Gaze behavioural differences are also particularly important for those with a diagnosis of an autism spectrum disorder (ASD). Overall attentional differences have been observed in ASD individuals, with less time being spent looking at the face and at the key face regions; eyes and mouth (see Chita-Tegmark, 2016 for a review; Grossman, Steinhart, Mitchell & McIlvane, 2015; Tanaka & Sung, 2016).

Atypical gaze patterns have been found for speech perception in noise for those with an ASD. For example, Irwin, and Brancazio (2014) found that children with ASD spent less time looking at the face of the talker and less time looking at the mouth area than typically developing children. This difference was not found in non-face and non-speech tasks, suggesting a speech specific gaze pattern. Irwin and Brancazio (2014) argue this leads to a reduction in access to speech information which may contribute to the communication deficits shown by typical ASD children. Conversely, using audio-stimuli taken from a feature film, Klin, Jones, Schultz, Volkmar and Cohen (2002) found ASD individuals spent significantly more time looking at the mouth region and significantly less time looking at the eye region than the control group. Klin et al. (2002) argue this reflects the nature of the task; complex social stimuli was used at the stimuli. The social content of the stimuli compels ASD individuals to search for meaning atypically, by avoiding the eye area which contains much of the social meaning. Avoidance of the eye area has been shown to be a typical trait of an individuals with ASD (Baron-Cohen, Wheelwright, & Jolliffe, 1997).

In a research study assessing gaze behaviour in response to audio-visual synchrony, ASD individuals were also shown to exhibit atypical responses. Grossman, Steinhart, Mitchell and McIlvane (2015) measured gaze behaviour while participants were looking at synchronous and asynchrous clear speech. ASD individuals looked at the synchrous stimuli less than controls and did not increase their gaze duration when explicitly asked to focus on the person speaking. Additionally, the ASD individuals spent significantly less time looking at the mouth area and significantly more time looking at the non-face regions on the screen. Grossman et al. (2015) argue this suggests individuals with ASD don’t maximise the opportunity to synchronise speech and integrate visual speech cues to aid speech perception.

Overall, this body of research suggests individuals with ASD respond atypically to speech perception tasks. This atypical response to speech perception has also been demonstrated in individuals who higher levels of autistic traits but who do not have a diagnosis of an ASD, although results have shown some variation.

Chen and Yoon (2011) measured aspects of autistic traits using sections from the Autism Spectrum Quotient (AQ). They used a general population sample and measured the association between scores on the AQ and gaze behaviour. They used a pre-recorded talking face in clear speech as the stimuli, with either a direct or averted eye gaze. Participants with a lower score on the AQ demonstrated an increased tendency to focus on the direct gaze than the averted gaze stimuli. Participants with a higher AQ score did not show this tendency. This may reflect reduced gaze reciprocity in those with higher self-reported autistic traits as measured by the AQ. Von dem Hagen and Bright (2017) also found differential gaze patterns that were dependent on the type of stimuli presented. In a video of a talking face, there was no difference in gaze patterns between individuals with higher and lower autistic traits. However, in a social interaction via a live video feed, higher autistic trait participants spent significantly less time looking at the conversation partner than the low autistic trait individuals. They argue the social presence of another individual significantly effects the gaze behaviour of participants with higher levels of autistic traits. In direct contrast, Freeth, Foulsham and Kingstone (2013) found reduced looking time towards the experimenter when using pre-recorded video stimuli than in the live interaction condition. However, as they point out, different participants were used in each condition (pre-recorded and live) which may have influenced the results.

Similarly, Vabalas and Freeth (2016), found no association between the levels of autistic traits in individuals and difference in looking time when engaged in a real-time social conversation with an experimenter. They found no difference between overall looking time or looking towards the face. However, they did find participants with higher levels of autistic traits had significantly less visual exploration of the face; measured by the number of and duration of eye saccades.

It could be argued that these contrasting results highlight the confounding effects of using different stimuli within research. As outlined in Experiment 3 in this project, there are differences in intelligibility and levels of visual speech benefit gained between different target talkers. It is possible differential gaze patterns reflect the relative audio or visual intelligibility of target talkers. There is currently a lack of research examining variation in gaze behaviour if the target talker changes and any association with the levels of autistic traits in participants.

In order to assess the gaze behaviour of participants, the audio-visual conditions in Experiment 5 were recorded using eye tracking. This was designed to examine gaze direction and duration. Research has shown that gaze is directed at the mouth area in challenging speech perception tasks, such as in background noise (e.g. Buchan et al. 2008; Yi, Wong and Eizenman, 2013). There is very little research examining gaze difference in clear and vocoded speech, but due to the challenging nature of the speech perception task, it was expected than gaze should be increasingly directed at the mouth area for both clear and vocoded speech. It was expected that increasing gaze at the mouth area would lead to significantly more visual speech benefit (Alsius et al., 2016; Rennig et al., 2018).

7.2.3 Hypotheses for Experiment 5

It was expected that clear speech would be easier to understand than vocoded speech and that audio-visual performance would be better than audio only performance. This is line with Experiments 1,2, 3 and 4 in this project. It was also hypothesised, that as found in Experiment 3, there would be more visual-speech benefit in vocoded speech than in clear speech for this target talker. It was expected that the familiarisation task would stabilise performance in the vocoded speech tasks, resulting in greater visual speech benefit in vocoded speech than in clear speech.

The following effects were hypothesised for clear and vocoded speech in terms of the regression model;

- For clear speech it was expected that general speech perception performance, ability to detect AV synchrony and a lower score on the autism spectrum quotient (AQ) would lead to significantly higher levels of visual speech benefit, as found in Experiment 4.
- For vocoded speech, it was expected that the familiarisation task would lead to less variation in performance in the vocoded speech tasks and therefore a significant regression model would be found. General speech perception performance, ability to detect AV synchrony and a lower score on the autism spectrum quotient (AQ) would lead to significantly higher levels of visual speech benefit.

- For the eye tracking data, it was hypothesised that significantly more time would be spent looking at the mouth area than the eye area for clear speech and for vocoded speech.
- For clear and vocoded speech, increased time looking at the mouth area, general speech perception performance, ability to detect AV synchrony and a lower score on the autism spectrum quotient (AQ) would lead to significantly higher levels of visual speech benefit for clear speech and for vocoded speech.

7.3 Method

7.3.1 Design

A within-participants design was used in Experiment 5. Several variables were measured; Audio only and Audio-visual Speech Reception Thresholds(SRT_{50}), measured in decibels (dB) were obtained for clear and vocoded speech. Further variables were overall performance in dB on the AV and AO tasks, score on the AV synchrony task and score on the AQ. In addition, the eye tracking measures were gaze direction and duration measured as the percentage looking time at each area of interest; the mouth and eyes.

The first analysis was a 2 (modality: audio only vs audio-visual) x 2 (speech type: clear vs vocoded) ANOVA. The dependent variable was SRTs measured in dB.

Secondly, a regression analysis was carried out. Clear and vocoded speech were analysed separately. The outcome variables were the amount of visual speech benefit gained (AV-AO SRTs, in dB for clear or vocoded speech). The predictor variables were general speech perception performance in dB (mean overall performance on the AO and AV tasks in clear speech and vocoded speech), percentage scored correctly on the audio-visual synchrony task, and score on the autism spectrum quotient (AQ).

The final analysis was carried out with the eye tracking data. Clear and vocoded speech were analysed separately. The first analysis compared the percentage dwell time on the mouth area and the eye area for clear and for vocoded speech. Secondly, a regression analysis for clear speech and for vocoded speech was carried out. The outcome variables were the amount of visual speech benefit gained (AV-AO SRTs, in dB for clear and vocoded speech). The predictor variables were general speech perception performance (mean overall performance on the AO and AV tasks in clear speech and vocoded speech), percentage scored correctly on the audio-visual synchrony task, score on the autism spectrum quotient (AQ) and percentage dwell time on the mouth area. Finally, additional analysis assessed dwell time on the mouth area as a function of the SNR's and as a function of accuracy.

7.3.2 Participants

The number of participants recruited for Experiment 5 was based on the criteria outlined in Experiment 1. Twenty-eight participants took part in this experiment (age 18-35, mean age 22, 3 males). Participants were recruited from the student population at Nottingham Trent University and were rewarded with research credits. Consent was obtained from each participant as agreed by the Nottingham Trent University Research Ethics Committee. Participants also confirmed normal hearing, normal or corrected to normal vision and had English as their first language. Two participants were excluded from the speech perception results because of incomplete data. Eight participants were excluded from the eye tracking results for clear speech and five participants from the eye tracking results for vocoded speech because of failure of the eye tracker to pick up the complete signal for all of AV sentences.

7.3.3 Materials and Procedure

7.3.3.1 Equipment

The experiments were conducted in a multi person IAC Acoustics 40a-5 audiology booth situated at Nottingham Trent University, as used in Experiment 2. The stimuli were presented in an identical manner to Experiment 2, with the exception of the screen size being 29.5cm x 47cm. The visual stimuli were presented at the identical size to that presented in Experiment 2, 17cm x 30cm. Eye movements were tracked using a SMI RED500 screen-based eye tracker which was also situated in the audiology booth.

7.3.3.2 Target Stimuli

The procedure for the practice task and the main experiment for this experiment were identical to that presented in Experiment 2. The sentences were chosen from the IEEE corpus (Rothauser et al., 1969). The same 80 sentences and the same target talker used in Experiment 4 were used for the main experiment. A further 30 sentences were recorded for the familiarisation task. A practice session was also included before the speech perception task, this duplicated Experiment 3.

7.3.3.3 Background Stimuli

Background noise for use in the experiment was created using the UCL speaker database (Markham & Hazan, 2002). Sixteen talkers (8 male and 8 female) were used as the background noise. This was the same background noise file used in Experiment 3.

7.3.3.4 Speech Processing

The vocoded speech was processed in an identical way to the previous experiments in this project (see section 4.3.3.4 for a full explanation of the procedure used).

7.3.3.5 Familiarisation Task

Participants initially undertook a familiarisation task using E-Prime software (Version 2.0 Psychology Software Tools Inc., Sharpsburg, US) for the vocoded stimuli. Thirty IEEE sentences were presented in a random order. They were presented at a range of SNR's; 2dB, 6dB and 10dB, with ten sentences being presented at each SNR. The SNR's were chosen to allow participants to experience a range of SNR's situated well within the range of performance levels established in Experiment 4. The sentences were processed using Matlab software in the same way as Experiment 4. They were presented in the same sixteen-talker background noise used in Experiment 4 and in the main experiment. The vocoded audio file for each sentence was presented to the participant, the sentence was then presented in text in the centre of the monitor screen for three seconds (size 40, courier new font, bold), and then the audio file was repeated (see Figure 7.1). Participants were instructed to watch and listen to the stimuli to aid familiarisation with the vocoded stimuli.

Figure 7.1. Example picture of the stimuli used in the familiarisation task.

7.3.3.6 General Speech Perception Ability

General speech perception ability was measured using the target stimuli as outlined in 6.3.3.2. Audio only and Audio-visual Speech Reception Thresholds(SRT_{50}), measured in decibels (dB) were obtained for clear and vocoded speech. The average of these were calculated for clear speech and for vocoded speech for each participant.

7.3.3.7 Audio-Visual Synchrony (AVS)

The audio visual synchrony task was presented in the same way as Experiment 4, with the following exceptions. Stimuli were presented in two blocks of 30 words, rather than one block of 60 words, as presented in Experiment 4. This was due to technical difficulties with the E-Prime software. The words were randomly allocated to either block one or two initially and remained in those blocks for all participants. The 30 words in each block was presented in a random order. The stimuli were presented at a size of 24cm high x 47cm wide on an HP monitor measuring 29.5cm x 47cm. The results were calculated in the same way as Experiment 4 and scored as percentage correct.

7.3.3.8 Autism Spectrum Quotient (AQ)

The AQ questionnaire (Baron-Cohen et al., 2001) was administered in the same way as Experiment 4. Participants were instructed to read each question carefully and circle how strongly they agree or disagree with each statement by circling the answer. Scores were totalled, giving an AQ score for each participant out of possible maximum of 50. A higher score on the AQ indicates greater self-reported levels of autistic traits.

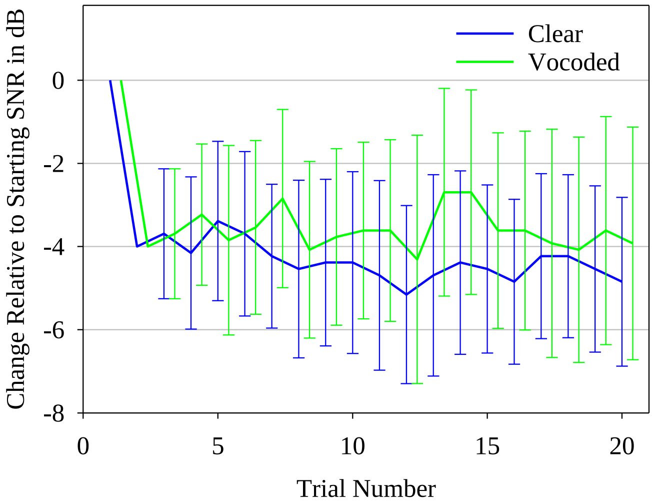
7.3.3.9 Eye Tracking

For the two audio-visual speech perception tasks, eye movements were tracked using a SMI RED500 screen-based eye tracker. The infrared light source and camera were situated below the screen allowing free head movement by the participants. Participants sat between 60-80cm from the monitor allowing comfortable viewing of the stimuli within the recommended distance from the eye tracker. The participants first completed a calibration task to ensure the eye tracker was correctly picking up their eye movements. This calibration process was repeated if necessary to ensure a satisfactory calibration level was reached. The calibration process was repeated before each of the AV tasks, once for clear speech and again for vocoded speech. The task then proceeded in an identical way to Experiment 4, with the addition of the eye tracker recording eye movements for each sentence presented in the AV conditions.

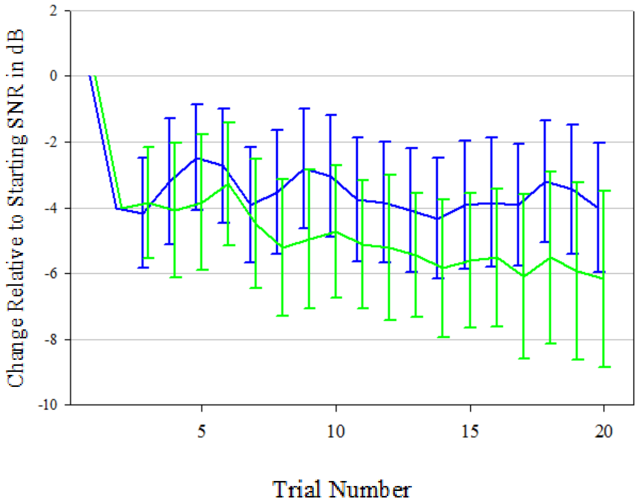
7.4 Results

7.4.1 Performance Over Time

The familiarisation task included in Experiment 5 was designed to enable participants to gain greater experience of listening to vocoded stimuli. Change over time for each participant over the 20 audio only sentences was calculated in the same way as Experiment 4. This was calculated by subtracting each participant's starting SNR from the SNR of each of the subsequent 20 sentences. This gave a value representing how far the SNR has moved from its starting value for each sentence. These values were averaged for each of the 20 sentences, giving a mean value per sentence. Figure 7.2 shows the change relative to the starting SNR for each participant for clear and vocoded speech for Experiment 5, and for comparison, the same data from Experiment 4.



Experiment 5



Experiment 4

Figure 7.2 The change relative to the starting SNR for each participant over the 20 sentences in the clear speech and vocoded speech audio only conditions for Experiment 5 and Experiment 4. Error bars show 95% confidence intervals.

Figure 7.2 shows, that in contrast to Experiment 4, participants in the vocoded speech condition in Experiment 5 did not move further away from their starting SNR in vocoded speech than in clear speech. The mean SNR for the final sentence for each participant was, on average, 4.9dB away from the starting SNR for clear speech and was 3.9dB for vocoded speech. This is in contrast to Experiment 4, where people moved further from their starting SNR on average for vocoded speech (6.2dB) than in clear speech (4dB).

In addition, the error bars for clear and vocoded speech are more similar in Experiment 5 than in Experiment 4. In Experiment 4 the error bars for the vocoded speech condition were wider than the clear speech condition suggesting that there was more variability in where participants ended the trials in comparison to the clear speech condition. This shows that performance in Experiment 5 for vocoded speech was more similar to clear speech.

7.4.2 Audio and Audio-Visual Performance

SRT_{50s} were calculated in the same way as Experiment 2, by averaging the last ten trials in each participant's adaptive track. All adaptive tracks were successful, with performance being between 30% and 70% correct as expected. There were 26 participants in the clear speech condition and in the vocoded speech condition. Figure 7.3 shows the average performance in each of the AO and AV tasks for clear and vocoded speech.

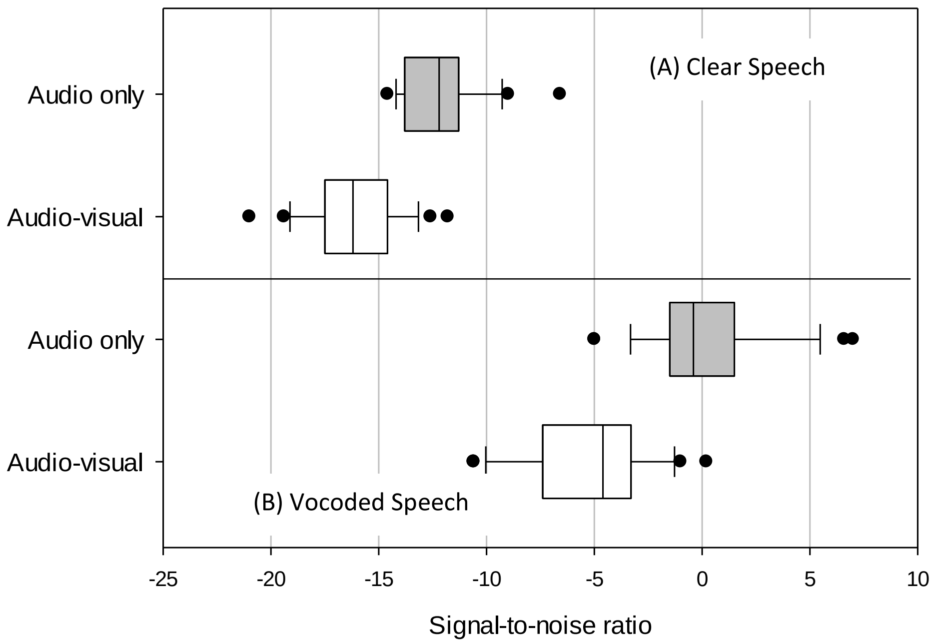


Figure 7.3. Audio only and audio-visual Speech Reception Thresholds for Clear (Panel A) and Vocoded (Panel B). The whiskers show the 10-90% range and the black dots show outlier data which fall outside that range.

A 2 (modality: audio only vs audio-visual) x 2 (speech type: clear vs vocoded) ANOVA showed a significant main effect of modality ($F(1,25) = 144.02$, $p<0.001$, $\eta_p^2 = 0.85$). Performance was better in audio-visual than audio only conditions. A significant main effect of speech type ($F(1,25) = 876.23$, $p<0.001$, $\eta_p^2 = 0.97$) was also found. Clear speech was easier to understand than vocoded speech. There was no interaction between modality and speech type ($F(1,25) = 3.09$, $p=.09$, $\eta_p^2 = 11$).

7.4.3 Effect of the Familiarisation Task on Vocoded Speech Performance

Experiment 5 included a familiarisation task to reduce variability in the vocoded speech conditions. In the AO vocoded speech condition, performance was similar for Experiments 4 (-0.26dB) and Experiment 5 (0.29dB). However, performance on the AV vocoded speech task was improved in Experiment 5 (M=-5.26dB, SD=3.06) over Experiment 4 (M=-3.78dB, SD=3.02) but not significantly ($t(49)=1.74$, $p=.83$). Similarly, there was greater visual speech benefit in Experiment 5 (5.52dB, SD=3.45) than in Experiment 4 (3.71dB, SD=3.53) but this was not significantly greater ($t(49)=1.85$, $p=.07$). The amount of visual speech benefit in Experiment 5 was also similar to that achieved in Experiment 3 (M=5.0, SD=3.4) where no familiarisation task was included. This suggests that although performance was less varied in Experiment 5, as a result of the familiarisation task, this did not significantly improve performance.

7.4.4 Visual Speech Benefit

The benefit received from the visual speech information was calculated by measuring the difference between performance levels in the AV and AO conditions. For clear speech, the mean visual speech benefit was 4.15dB (SD=2.38) and for vocoded speech was 5.52dB (SD=3.45). There was no significant difference between the amount of visual speech benefit received between clear and vocoded speech ($t(25)=1.67$, $p=.11$).

The amount of visual speech benefit for the clear speech condition and the vocoded speech condition are shown in Figure 7.4. As found in Experiment 4, the variation in the amount of visual speech benefit was greater in the vocoded speech condition than the clear speech condition, shown by the wider distribution of the results.

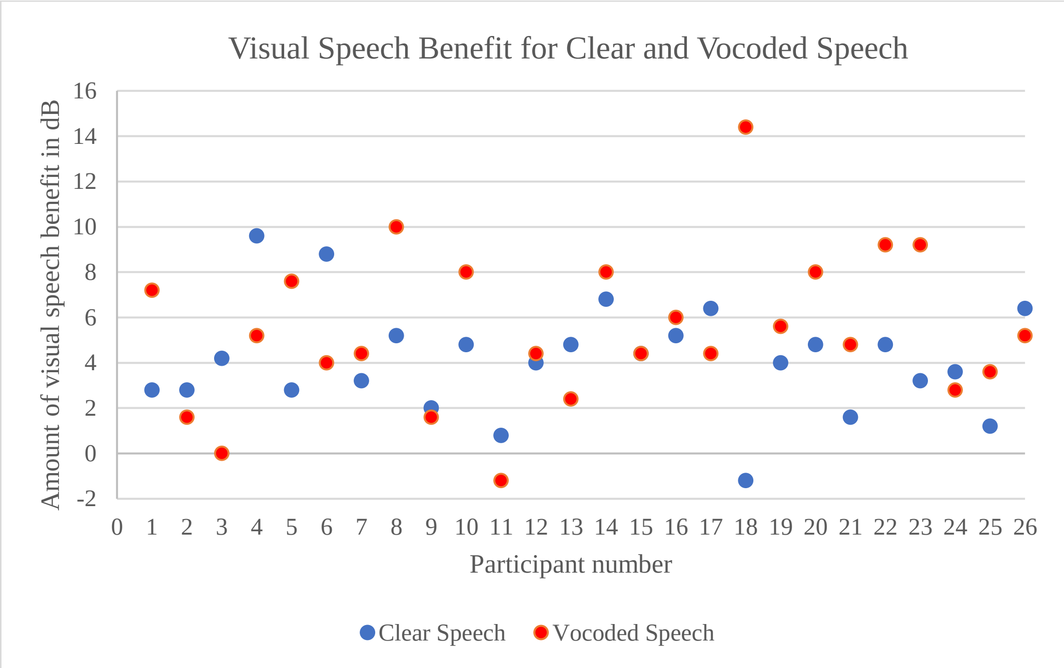


Figure 7.4 The distribution of the amount of visual speech benefit received for clear and vocoded speech.

7.4.5 Audio-Visual Synchrony (AVS)

Results were scored as percentage correct. The scores ranged from 33.3% to 85% correct. The mean score was 61% correct (SD=10.1).

7.4.6 Autism Spectrum Quotient (AQ)

Scores on the AQ ranged from 3 to 26 (M=14.77, SD=6.22). These results are in line with expected scores whereby an average score is defined as being between 13 and 20 (von dem Hagen & Bright, 2017).

7.4.7 Regression Analysis

A regression analysis was run to ascertain if the levels of visual speech benefit received were predicted by average overall AV and AO performance, scores on the AV synchrony task and score on the AQ. All assumptions of the regression analysis were met; the data was normally distributed and had homogeneity of variance, no multicollinearity; no significant correlations between variables, tolerance and VIF were within acceptable levels.

Clear speech

For clear speech, the model was not significant. There were no significant predictors of the amount of visual speech benefit received, with 12% of the variance being explained (adj. R^2 11.9%, F=2.13, p=.13) (Table 7.1). The AV synchrony task had a positive relationship with visual speech benefit, better performance on the AV synchrony task led to greater visual speech benefit. The AQ score had a negative relationship with visual speech benefit; a higher score on the AQ led to lower levels of visual speech benefit. The AQ result was approaching significance (p=.083).

Table 7.1 Results of the regression analysis for clear speech

Variable	B	SE(B)	β	t	p
AVS	.07	.05	.28	1.47	.16
AQ	-.13	.07	-.34	-1.82	.083
Overall Performance	-.25	.28	-.17	-.89	.39

Vocoded Speech

For vocoded speech the model was not significant. There were no significant predictors of the amount of visual speech benefit received, with 0.7% of the variance explained (adj. R^2 .066%, F=1.59, p=.22) (Table 7.2).

This result was similar to that found in Experiment 4 and therefore not affected by the inclusion of the familiarisation task for vocoded speech, which was designed to reduce variability in performance.

Table 7.2 Results of the regression analysis for vocoded speech

Variable	B	SE(B)	β	t	p
AVS	-.08	.07	-.28	-1.14	.27
AQ	.19	.11	.34	1.77	.092
Overall Performance	-.22	.27	-.16	-.82	.42

7.4.8 Eye Tracking Analysis

Eye tracking data from the SMI RED500 was analysed using SMI software (Begaze version 3.7 build 42). Conditions where the eye tracker failed to pick up gaze information for all of the sentences were excluded from analysis. This resulted in full data sets remaining for 20 participants in the clear speech condition and 23 participants in the vocoded speech condition. The remaining participants were different for the clear and vocoded speech conditions and therefore comparisons across clear and vocoded speech types are not made in this analysis.

Areas of interest were created on each of the stimuli; the mouth area and the eye area. This is illustrated in Figure 7.5, which shows an example of the areas of interest drawn for each of the stimuli. The areas of interest were made large enough to accommodate the mouth when fully opened.

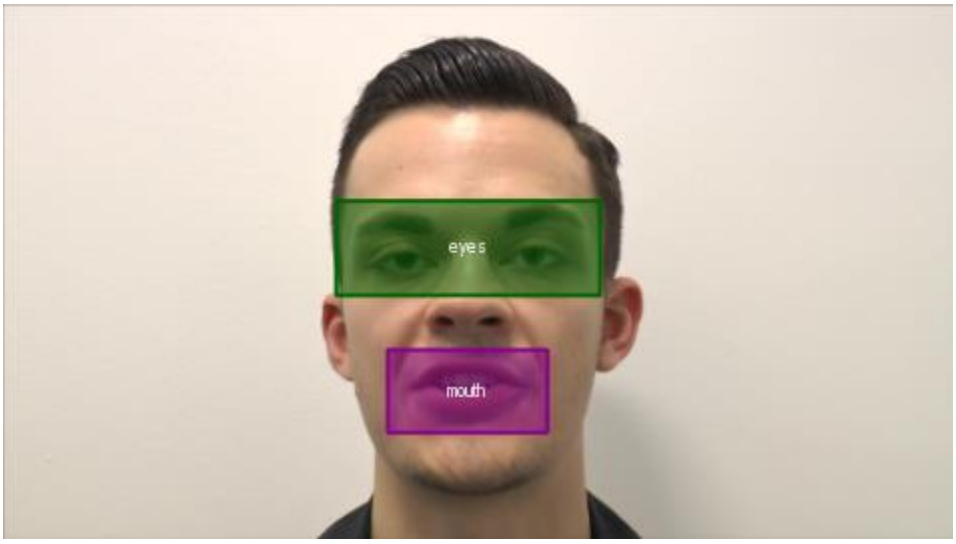


Figure 7.5. An example of the stimuli used for each sentence with the AOI's for the mouth and eye areas shown.

The dwell time for each sentence for each participant was calculated for each of the areas of interest. The first two sentences in each condition were excluded from the analysis because the eye movements recorded for these two sentences showed the repetition of the initial sentence to establish the point at which 50% of the key words could be identified, combined with the second sentence in each condition. Therefore, these first two sentences did not contain eye tracking data for a single individual sentence and were excluded from the analysis. This left 18 sentences for each condition for analysis. The mean percentage dwell time for each area of interest are shown in Table 7.3.

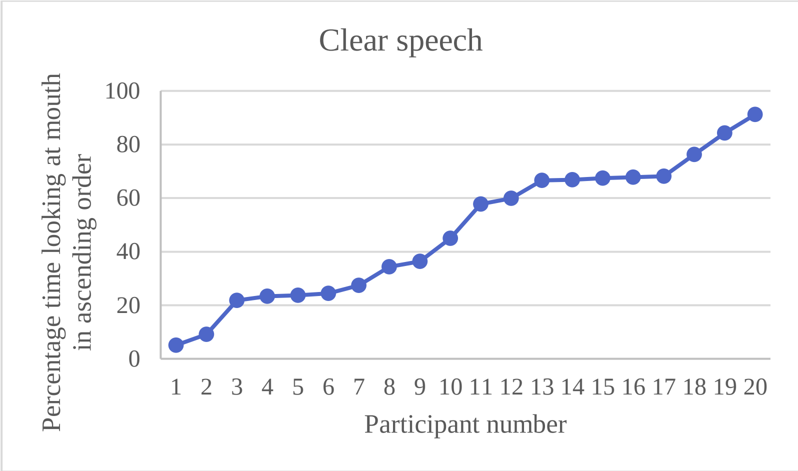
Table 7.3. Mean percentage dwell time on each area of interest for clear speech and vocoded speech

	Clear speech Mean	SD	Vocoded Speech Mean	SD
Mouth	47.9	24	46.8	21.1
Eyes	6.2	5.1	9.2	6.9

These results show significantly more time was spent looking at the mouth area than the eye area for clear speech ($t(19)=8$, $p<.001$) and for vocoded speech ($t(22)=8.4$, $p<.001$).

Variation between participants in looking time

There was a considerable variation in the percentage of time spent looking at the mouth area between participants, as demonstrated by the large standard deviation in the mean dwell time at the mouth area (M=47.9, SD=24.0 for clear speech, M=46.8, SD=21.1 for vocoded speech) (Table 7.3). Figure 7.6 shows the percentage dwell time on the mouth area in ascending order, from least amount of time spent looking at the mouth area, to the highest percentage of dwell time on the mouth area. Data is shown for clear speech and for vocoded speech.



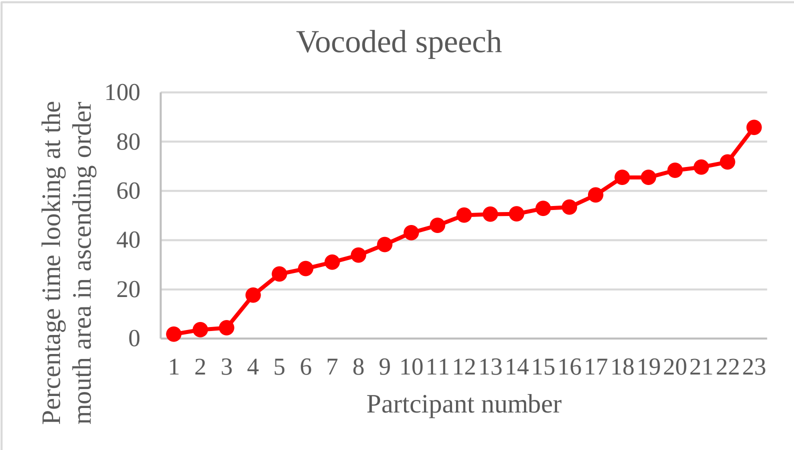


Figure 7.6 The percentage of time spent looking at the mouth area for each participant in ascending order for clear speech and for vocoded speech.

Variation in Gaze Patterns

There was also considerable variation between participants in gaze behaviour recorded by the eye tracker. For example, for clear speech, the lowest percentage of time spent looking at the mouth area was 5.07%. Figure 7.7 shows the scan path and heat map for a randomly chosen sentence for the participant with the lowest amount of time spent looking at the mouth area. The stimuli sentences last for approximately three to four seconds each. The circles in the scan path diagram represents the amount of time spent looking at this area, a larger circle shows more time was spent looking at this area.

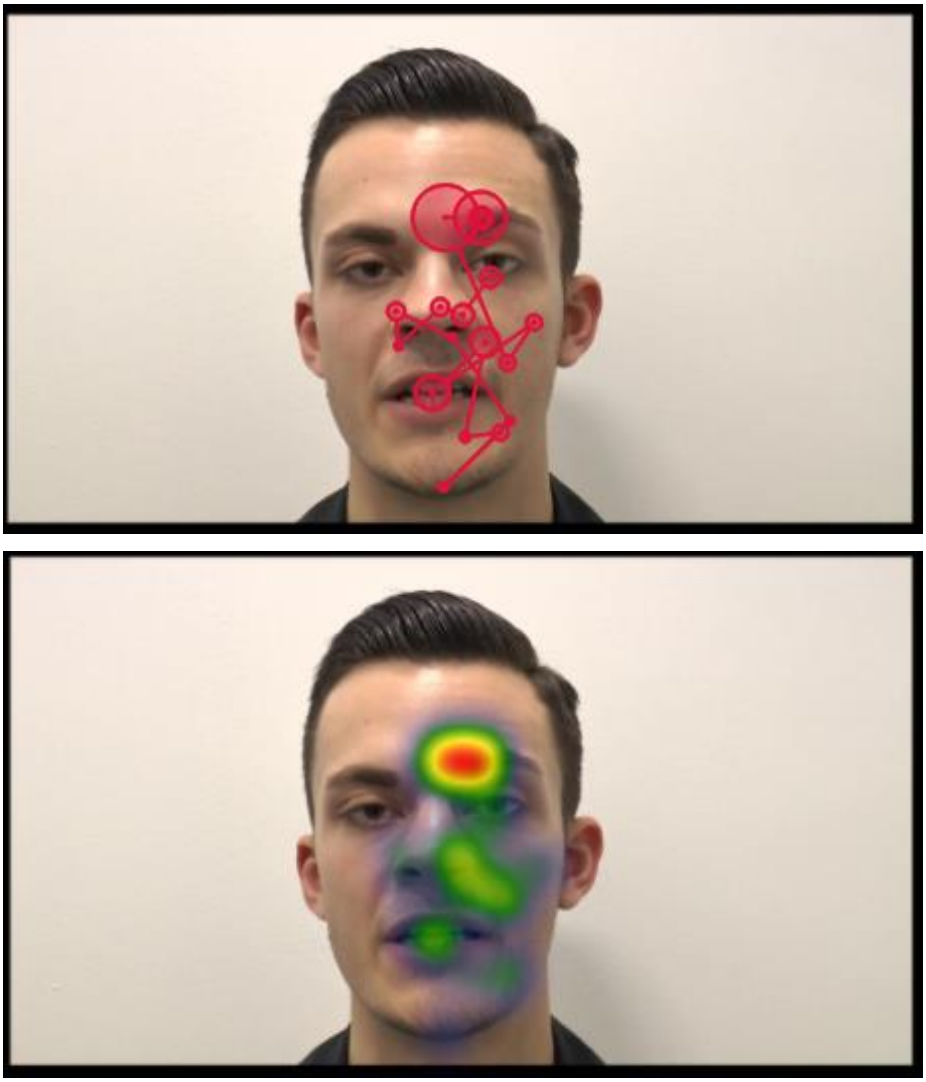


Figure 7.7 This shows a scan path and heat diagram for a participant with a small percentage of time spent looking at the mouth area. These diagrams are for one sentence in clear speech.

Figure 7.7 contrast with Figure 7.8 which show the same data for a participant with a higher percentage of time spent looking at the mouth area. This participant spent 84.28% of the time overall looking at the mouth area. Figure 7.8 shows the scan path and heat diagram for one sentence for this participant in clear speech. It clearly shows a larger percentage of time looking at the mouth area, shown by the larger circles in the scan path diagram. The heat map also shows an increased focus on the mouth area; the red colour shows the greatest area of focus through yellow to green and finally to the blue colour, which shows the least amount of focus on that area.

The scan path diagram in Figure 7.7 also shows considerable shift in gaze by the participant with the lowest amount of dwell time on the mouth area. The lines joining the circles represent the eye movements around the stimuli. This shows extensive gaze shift around the face within the three to four seconds the stimuli was being played to the participant. This again contrasts with Figure 7.8, which shows a participant with very high dwell time on the mouth area, where the gaze focuses mainly on the mouth area for a longer time span and does not move around the face.

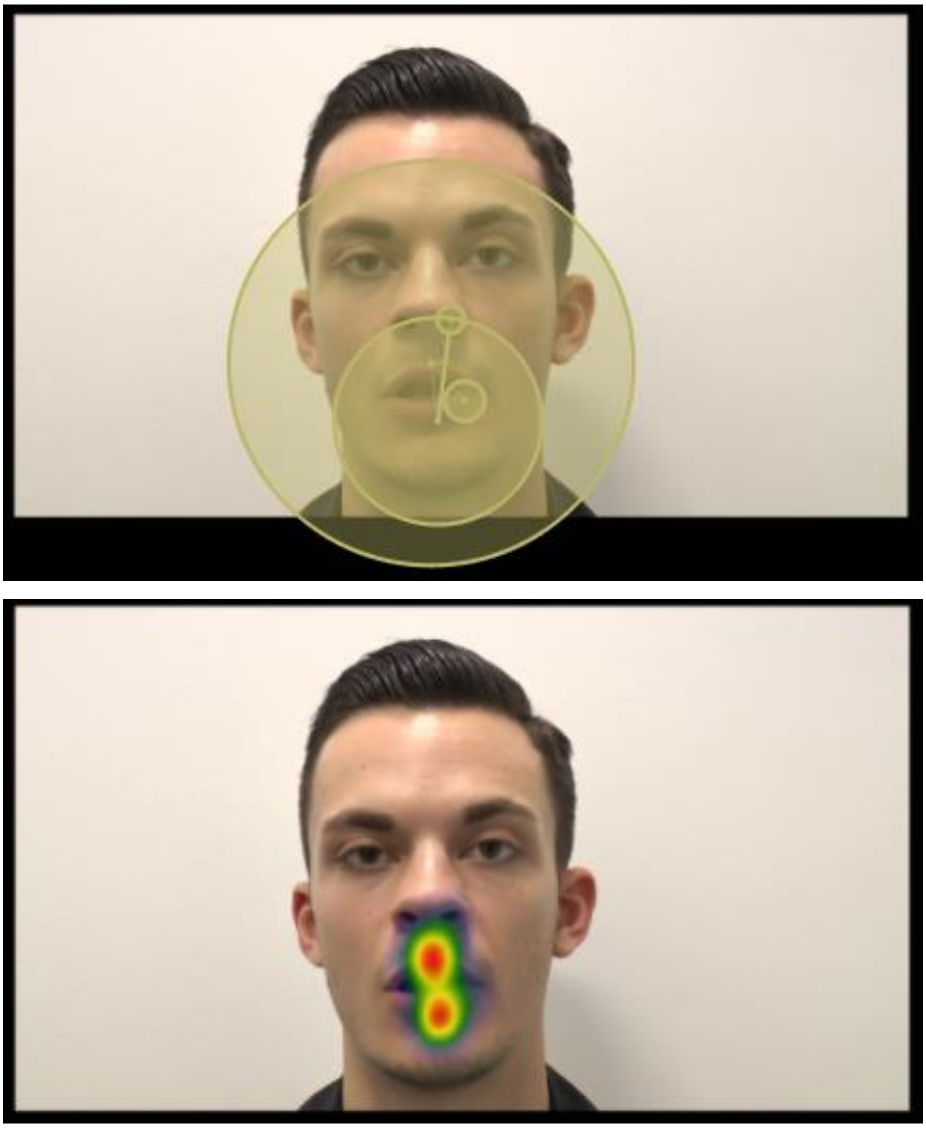


Figure 7.8 Showing a scan path and heat diagram for a participant with a high percentage of time spent looking at the mouth area. These diagrams are for one sentence in clear speech.

7.4.9 Regression Analysis with the Eye Tracking Data

Clear Speech

The regression analysis was re-run on the 20 remaining participants in the clear speech condition to include the percentage dwell time on the mouth. This tested whether the amount of visual speech benefit received was predicted by the average overall performance in dB on the AV and AO tasks, score on the AV synchrony task, score on the AQ and the mean percentage dwell time on the mouth area. The mean overall performance for the remaining 20 participants in shown in Table 7.4.

Table 7.4 Mean results for the predictor variables in clear speech

Variable	Mean	SD
AVS	62.4	10.8
AQ	15.4	6.1
Overall Performance	-14.3dB	1.7
Mean % Mouth Area	47.6	25.7

The results of the regression show that the model was significant, with two of the predictors explaining 50.4% of the variance (adj. R^2=.504, F(4,15)=5.82, p=.005). The average overall performance significantly predicted the amount of visual speech benefit received (β=-.59, p=.004), suggesting better overall performance on the task resulted in increased visual speech benefit. The second significant predictor was the mean percentage time spent looking at the mouth area (β=.42, p=.029), suggesting that more time spent looking at the mouth area resulted in increased visual speech benefit. Table 7.5 shows the regression analysis.

Table 7.5 Results of the regression analysis for clear speech

Variable	B	SE(B)	β	t	p
AVS	.05	.03	.27	1.64	.12
AQ	-.06	.06	-.18	-.98	.34
Overall Performance	-.74	.22	-.59	-3.39	.004
Mean % Mouth Area	.04	.02	.42	2.41	.029

This regression model, including the eye tracking data, is now significant as shown in Table 7.5. This is in contrast to the original regression model, including all 26 participants, where the overall model was not significant (adj. R^2 11.9%, F=2.13, p=.13) (Table 7.1).

In order to establish if this significance was driven by the eye tracking data or by the change to the participant sample, the regression mode was re-run without the eye tracking data (mean percentage dwell time on the mouth area) for the 20 participants used in the eye tracking model. This model was significant, with one of the variables explaining 35.4% of the variance (adj. R^2=.35, F(3,16)=4.48, p=.018). The average overall performance significantly predicted the amount of visual speech benefit received (β=-.48, p=.024), suggesting better overall performance on the task resulted in increased visual speech benefit. Table 7.6 shows the revised regression analysis without the eye tracking data.

Table 7.6 The revised regression analysis without the eye tracking data.

Variable	B	SE(B)	β	t	p
AVS	.05	.04	.27	1.41	.18
AQ	-.11	.07	-.33	-1.7	.11
Overall Performance	-.6	.24	-.48	-2.5	.024

Vocoded Speech

The regression analysis that was carried out on the clear speech data outlined above, was repeated for the vocoded speech data. The overall mean performance for the remaining 23 participants in the vocoded condition is shown in Table 7.7.

Table 7.7 Mean results for the predictor variables in vocoded speech

Variable	Mean	SD
AVS	60.7	9.7
AQ	15.1	6.2
Overall Performance	-3.1	2.6
Mean % Mouth Area	44.2	23.1

The results of this analysis showed that for vocoded speech there were no significant predictors of the amount for visual speech benefit received (adj. R^2=.023, F(4,18)=1.13, p=.37). Table 7.8 shows the regression analysis results for each of the predictor variables.

Table 7.8 Results of the regression analysis for vocoded speech

Variable	B	SE(B)	β	t	p
AVS	-.08	.07	-.24	-1.01	.3
AQ	.17	.11	.34	1.57	.13
Overall Performance	-.15	.27	-.13	-.56	.58
Mean % Mouth Area	.04	.03	.3	1.38	.19

7.4.10 Additional Eye Tracking Analyses

Dwell time on the mouth area as a function of SNR

The median SNR for each participant was calculated over the 18 sentences to establish if gaze duration at the mouth area changed as a function of a change in the SNR. The sentences were divided into two groups based on the median SNR for each participant; high and low SNR. The mean percentage dwell time on the mouth area was calculated for each of these groups. Any sentences that were equal to the median SNR were disregarded.

For clear speech, there was no difference between the percentage of time spent looking at the mouth area between higher SNR sentences ($M=50.21$, $SD=24.3$) and lower SNR sentences ($M=47.61$, $SD=23.2$; $t(19)=1.1$, $p=.28$). Similarly, for vocoded speech, there was no difference between the percentage of time looking at the mouth area for higher SNR sentences ($M=46.86$, $SD=22.3$) and lower SNR sentences ($M=48.17$, $SD=24.79$; $t(22)=-3.77$, $p=.71$). This demonstrates that the gaze duration on the mouth area did not change if the SNR was increased or decreased.

Dwell time on the mouth area as a function of accuracy

The 18 sentences were divided into accurate and inaccurate groups for each participant; accurate where three or more of the key words were correctly identified and inaccurate where between zero and two words were correctly identified. The mean percentage dwell time on the mouth area was calculated for each participant for the inaccurate and accurate sentences. The overall mean percentage dwell time for the accurate and inaccurate sentences was then compared to establish if this varied if the sentence was accurately or inaccurately identified.

For clear speech, there was no difference between the percentage of time spent looking at the mouth area for inaccurate sentences ($M=49.32$, $SD=23.3$) and accurate sentences ($M=47.25$, $SD=23.8$; $t(19)=1.3$, $p=.20$). For vocoded speech, there was also no difference between the percentage of time spent looking at the mouth area between inaccurate ($M=47.71$, $SD=20.4$) and accurate sentences ($M=46.01$, $SD=22.6$; $t(22)=.79$, $p=.44$). This illustrates that the amount of time spent looking at the mouth area did not change if the participants were able to successfully or unsuccessfully identify three or more of the key words for clear or vocoded speech.

Dwell time on the mouth area as a function of a high or low score on the AQ

Participants were grouped into high and low AQ scoring groups; high if a score of more than 20 was achieved on the AQ, and low if a score of less than 13 was achieved (von dem Hagen & Bright, 2017). Five participants were grouped into the high AQ group (M=23, SD=2.24) and six participants into the low AQ score group (M=8, SD=3.03). Percentage dwell time on the mouth area for each group (High AQ M=34%, SD=30.45; low AQ M=58.24, SD=58.24, SD=18.48) was compared and showed no significant difference between the groups ($t(9)=1.6$, $p=.14$).

7.5 Discussion

7.5.1 Summary of Overall Results

The results of Experiment 5 suggest that as expected, clear speech was easier to understand than vocoded speech and that audio-visual speech was easier to understand than audio only speech. This is in line with previous experiments in this project. There was no difference in the amount of visual speech benefit in clear and vocoded speech, as found in Experiment 4. The inclusion of the familiarisation task reduced variability in the vocoded speech conditions but did not improve overall performance levels. The overall regression model was not significant for clear speech or for vocoded speech. However, when the eye tracking data was added to the model for clear speech (time spent looking at the mouth area), the model was significant. General speech performance and time spent looking at the mouth area explained 50% of the variance in performance levels. This model was reanalysed without the eye tracking data and remained significant, with general speech performance explaining 35% of the variance in performance levels. These results are discussed in detail in the next section.

7.5.2 Performance Over Time for Clear Speech and for Vocoded Speech

The results of Experiment 5 show that performance in the vocoded speech condition was less varied than in Experiment 4. In contrast to Experiment 4, the change in performance over the course of the 20 sentences for the audio only condition for vocoded speech (3.9dB) was less than the clear speech sentences (4.9dB). In Experiment 4, participants moved 6.2dB away from their starting SNR in vocoded speech, and 4dB in clear speech. This suggests the familiarisation task stabilised performance as expected; performance in clear speech and vocoded speech was more similar over the course of the 20 sentences. This is also shown by the width of the error bars. The error bars in Experiment 5 are more similar in clear and vocoded speech. This is in contrast to Experiment 4 where the error bars were wider for vocoded speech, suggesting increased variation across participants as the task progressed. Taken together, this suggests the familiarisation task was successful in providing enough experience with the vocoded stimuli to stabilise performance so that performance in the clear and vocoded tasks was more similar.

The success of the familiarisation task supports the research of Davis et al. (2005). Davis et al (2005) used 30 sentences as a training task and found significant improvement over the course of the sentences in identifying key words. Davis et al. (2005) used six channel noise vocoded speech. Davis et al. (2005) also argued that written presentation of the sentence provided the same amount of benefit to understanding of the vocoded speech as hearing the sentence in clear speech. Experiment 5 presented the sentences in the familiarisation exercise in written form, in an identical way to Davis et al. (2005). This suggests the ‘pop-out’ effect noted by Davis et al. (2005) may have assisted participants in Experiment 5 in gaining more familiarity with the vocoded stimuli. This allowed increased stability in performance and less improvement over time during the speech perception task was found.

Despite the reduction in variability in the vocoded speech tasks, there was no significant difference in performance in the audio only and audio-visual conditions between Experiment 5 and Experiment 4. This suggests the familiarisation task did not improve overall performance levels in the vocoded speech conditions. This is discussed in the next section.

7.5.3 Comparison Between Clear and Vocoded Speech

The amount of visual speech benefit was not significantly different in clear and vocoded speech. This in contrast the expected results but is in line with the results of Experiment 4. Additionally, the distribution in the amount of visual speech benefit was greater in the vocoded speech condition than the clear speech condition, which mirrors the results of Experiment 4. The familiarisation task was intended to address some of this variability and find some significant predictors of visual speech benefit for vocoded speech. While the familiarisation task was successful in removing the variability in change from the starting SNR for vocoded speech, the wider distribution in the amount of visual speech benefit was not affected by the inclusion of the familiarisation task. This suggests the variability in the amount of visual speech benefit gained by participants was not influenced by lack of experience with the vocoded stimuli in comparison to the clear speech stimuli. It could be argued that whilst the familiarisation task gave participants enough experience with the vocoded stimuli to address the change in performance over time (performance over time became more similar in vocoded speech and clear speech), there was not enough experience gained by some participants to enable a more similar performance in the clear speech and vocoded speech condition. Rosen et al, (1999) demonstrated that more sustained training with vocoded stimuli results in substantial gains in understanding of vocoded speech; improvements from 1% correctly identified keywords to nearer 40% correctly identified keywords after three hours of training. Therefore, further training and experience with vocoded stimuli could benefit some participants and result in less variability in the amount of visual speech benefit gained.

This greater variation in the amount of visual speech benefit gained in vocoded speech than in clear speech may also explain the lack of fit of the regression model for vocoded speech. It was expected that with more experience with the vocoded speech stimuli, the ‘super visual-speech perceiver’ identified in the clear speech model, would be found in the vocoded speech model. This was not found. This could suggest that the familiarisation task did not give participants enough experience with the vocoded speech to establish any expertise. More extensive training could be given to participants who gained greater visual speech benefits in clear speech to establish if this would change their performance in vocoded speech.

Alternatively, it may not be possible to achieve the same level of expertise with vocoded speech as clear speech, given that participants would have spent a life time gaining experience with clear speech. To address this, the experiment could be repeated with experienced cochlear-implant users, who are familiar with vocoded speech, to assess if their levels of visual speech benefit in noise are predicted by greater abilities on audio only and audio-visual speech perception, ability to detect AV synchrony and a lower score on the AQ. This would ascertain if ‘super visual-speech perceiver' expertise could be acquired through extensive experience with the most familiar type of speech to each participant.

7.5.4 Overall Model for Clear Speech

In contrast to Experiment 4, overall greater abilities on audio only and audio-visual speech perception, ability to detect AV synchrony and a lower score on the AQ were not predictive of the amount of visual speech benefit gained in clear speech. Two of the predictors were in the same direction as expected; the AV synchrony task had a positive relationship with visual speech benefit, better performance on the AV synchrony task led to greater visual speech benefit and the AQ score had a negative relationship with visual speech benefit; a higher score on the AQ led to lower levels of visual speech benefit. General speech perception performance had a marginally negative relationship with the amount of visual speech benefit gained, which was not as expected. The results of the audio only and the audio-visual speech tasks in Experiment 5 were similar to Experiment 4. However, the results of the other two predictor measured vary between Experiment 4 and 5. The AV synchrony task in Experiment 5 had significant lower score than in Experiment 4 (M=61%, SD=10.1 in Experiment 5, M=77%, SD=12.29 in Experiment 4). This may have been influenced by changes to the stimuli, whereby the task was run in two separate blocks in Experiment 5 and in one continuous block in Experiment 4. Additionally, the range of scores in the AQ test was considerably less in Experiment 5 than in Experiment 4. The AQ scores in Experiment 5 ranged from 3 to 26, while in Experiment 4 the AQ scores ranged from 4 to 35. The mean AQ score and standard deviation were also higher in Experiment 4 than in Experiment 5. This suggests greater variability in AQ scores in Experiment 4 than in Experiment 5. The relative lack of variability in the AQ scores in Experiment 5 and lower AV synchrony scores in Experiment 5 may account for differences in the significance of the model. It is also worth noting that the sample size used in this experiment is limited, so small differences in the AQ range and the AV synchrony scores are disproportionately effecting the results. This is discussed in further detail in Chapter 8 where comparisons of results across experiments is made.

7.5.5 Eye Tracking Results

The eye tracking shows that as hypothesised, more time was spent looking at the mouth area than the eye area. Although comparisons cannot be directly made across the speech types (the results were taken from different participants and separate analysis undertaken for the clear and vocoded groups), it is interesting to note than the amount of dwell time on the mouth area and eye area did not differ between the groups (Table 7.3). This suggests that task difficulty was equally challenging for participants in both speech types; the time spent looking at the mouth area was not driven by the speech type but perhaps by the challenging background noise. Vatikiotis-Bateson et al. (1998) found that 55% of the overall gaze time was directed at the mouth area for clear speech in a high level of babble type background noise. This is comparable to Experiment 5, where 47.9% of the dwell time was directed at the mouth area for clear speech in a 16 talker background noise. Although stimuli length differed, Vatikiotis-Bateson et al. (1998) used 35-45 second monologues, and Experiment 5 used IEEE sentences lasting 3-4 seconds, similar dwell times were found in both studies. This suggests similar strategies are being used where background noise is particularly challenging.

The variation in the amount of dwell time on the mouth area was also considerable. For clear speech the amount of mouth dwell time varied between 5.07% and 84.28% and for vocoded speech varied between 1.76% and 85.78%. This suggests very different perceptual strategies are being used by participants. This is also demonstrated in the examples of gaze patterns and heat maps for a participant with very low dwell time on the mouth area (Figure 7.7) and a higher percentage dwell time on the mouth area (Figure 7.8). Given that for clear speech, the regression analysis showed that the percentage dwell time looking at the mouth area was a significant predictor of the amount of visual speech benefit gained, it would suggest that focusing on the mouth area is the most effective strategy for gaining visual speech benefit in clear speech. The more time spent looking at the mouth area results in the most benefit from visual speech information. Alsius et al. (2016) found increasing levels of visual speech benefit for participants who spent more time looking at the mouth area for words and sentences in a multi-talker babble background noise with and SNR’s of -2dB for words and -3dB for sentences. In addition, Alsius et al. (2016) also report considerable variation between participants in the amount of visual speech benefit gained.

Rennig et al. (2018) found significantly more visual speech benefit was gained by participants who looked more at the mouth area when identifying single syllables in clear speech but not when using sentences in pink noise at -16dB. However, Rennig et al. (2018) found that 93% of the time on average was spent looking at the mouth area with low variability between participants in the sentences in -16dB pink noise. This suggests this task was too difficult for participants to complete without considerable attention being focused on the mouth area. This is in contrast to the results of Experiment 5 where considerable variation between looking time at the mouth area was found. Rennig et al. (2018) report considerable variation in the clear speech syllable task; between 6% and 56% of time looking at the mouth area. This would suggest gaze behaviour may be task dependent and vary as a result of changes to the type of background noise.

Taken together, this would suggest the most effective strategy in gaining visual speech benefit in increased focus on the mouth area. However, given the large variation between participants in the amount of visual speech benefit gained, and the eye tracking evidence from Experiment 5, this strategy is not being used by all participants. Therefore, other variables are influencing the strategies of participants when trying to understand visual speech.

The significant predictors of the amount of visual speech benefit in clear speech, for the 20 participants used in the eye tracking study, were the amount of time spent looking at the mouth area, as outlined above, and general speech perception performance. This regression analysis was rerun without the eye tracking data as a predictor variable, and overall speech perception performance remained a significant predictor of the amount of visual speech benefit gained. This suggests general speech perception performance remains a significant predictor of the amount of visual speech benefit gained, regardless of the gaze behaviour of the participants. Ability to detect audio-visual synchrony and score on the AQ were not predictive of the amount of visual speech benefit gained. Additionally, there was no difference in the amount of dwell time on the mouth area as a result of being grouped as a high or low AQ participant. Although not significant, the direction of the AV synchrony result and score on the AQ were as expected. This is discussed in further detail in Chapter 8 where comparisons across experiments is made.

The additional analysis using the eye tracking data showed that there was no change in the dwell time on the mouth area as a function of the SNR or as a function of accuracy. For lower and higher SNR's there was no significant difference in the amount of time spent looking at the mouth area. Similarly, there was no change in the amount of time spent looking at the mouth area if the keywords were correctly or incorrectly identified. This suggests gaze behaviour does not change as a result of changes to the SNR or accuracy, and therefore, may be driven by individual differences in participants. Gaze behaviour is being driven by the perceptual strategies used by participants. This is in contrast to Lusk and Mitchel (2016) who found gaze shifted away from the mouth area as the artificial langue they used became more familiar. This suggests the stimuli used in Experiment 5 were significantly challenging to engage participants gaze throughout the task, and no floor effects were reached where gaze shifted away from the most effective area of the stimuli that would aid speech perception.

Overall, the eye tracking results suggest particularly challenging background noise, as used in Experiment 5, leads to increased focus on the mouth area to aid visual speech perception. This is the most effective strategy to improve visual speech understanding. However, not all participants are using this strategy to aid visual speech perception and are therefore not able to gain as much visual speech benefit. It is possible that changes to the type of background noise, making it less challenging may change gaze behaviour. However, as gaze patterns did not change as a result of changes to the SNR and as a function of accuracy, it is also possible that gaze patterns are driven by individual differences between participants.

In order to investigate this further, Experiment 5 could be repeated using less challenging background noise types, for example, one or two talkers as used in Experiment 1 and 2. Gaze patterns could be assessed to see if they vary as a result of changes to the background noise, or if gaze patterns remain the same regardless of background noise types. Participants who use the most effective perceptual strategies for gaining the most visual speech benefit may continue to use this strategy regardless of the background noise type. Alternatively, gaze behaviour may change if the background noise type is less challenging. However, any type of background noise creates masking effects and makes speech understanding more challenging, as demonstrated in Experiments 1 and 2, so any type of background noise may provoke the same perceptual strategies.

It would also be interesting to repeat this procedure with explicit instructions given to participants to look at the mouth area of the target talker. Increased focus on the target talker should lead to increased levels of visual speech benefit. If participants are able to change their strategy and this results in higher levels of visual speech gain, it would be useful to advise people to focus on the mouth are to aid speech understanding, particularly in background noise. This may also be a useful strategy for hearing impaired populations to employ.

7.6 Chapter Summary

- Individual differences in the amount of visual speech benefit gained by participants was assessed for clear and vocoded speech
- The predictors that were assessed were general speech perception ability, ability to detect audio and visual synchrony, tendency towards autistic traits and gaze direction and duration measured by eye tracking
- For clear speech, two of the predictors explained 50% of the variance in the amount of visual speech benefit gained, general speech perception ability and time spent looking at the mouth area
- A familiarisation task was included to stabilise performance over time in the vocoded speech conditions. Performance over time was stabilised but performance levels did not improve in the vocoded speech conditions
- For vocoded speech, there were no significant predictors of the amount of visual speech benefit gained
- There was no difference in the amount of visual speech benefit gained between clear and vocoded speech.

Chapter 8 – Comparison of Results across Experiments

This chapter compares results across experiments in this project in order to gain further understanding of the results of this research. Firstly, comparison of the target talker used in Experiments 1 and 2 is made with the four different talkers used in Experiment 3. It is possible the target talker used in Experiments 1 and 2 had lesser or greater levels of intelligibility compared to the target talkers used in Experiment 3. The intelligibility of this talker may have affected the results of Experiment 1 and 2 as a result of the intelligibility level of this talker. Secondly, the results of Experiment 4 and 5 are examined with reference to the results for the clear speech conditions. The results of Experiment 5 did not demonstrate any significant predictors of the amount of visual speech benefit gained by participants for clear speech. However, the general direction of the data from Experiment 5 was similar to that of Experiment 4. It is therefore possible that a significant relationship may be found if the data from these two experiments is combined.

8.1 Intelligibility of the Talker used in Experiments 1 and 2

The target talker used in Experiment 1 and 2 was the same talker. Four different talkers were used in Experiment 3, where assessment of the intelligibility of talkers was examined. There was a significant difference in audio intelligibility across the four talkers in Experiment 3, with the most unintelligible talker in clear speech, providing the most visual speech benefit. The amount of visual speech benefit gained increased as the audio intelligibility of the target talker decreased in clear speech.

The relative intelligibility of the target talker used in Experiments 1 and 2 is shown in Figure 8.1. The figure shows that the target talker in Experiments 1 and 2 was relatively unintelligible when compared to the four talkers assessed in Experiment 3. For example, for the sixteen talker background noise, the average results in clear speech for the talker in Experiment 2 were -4dB in the audio only condition and -8.6dB in the audio-visual condition. This results in the talker from Experiments 1 and 2 being the second most unintelligible talker in both the audio and audio-visual speech condition for clear speech. Similarly, for vocoded speech, the average performance levels in the sixteen talker background noise were 6.9dB in the audio only condition and 3dB in the audio-visual condition. This again results in the talker used in Experiments 1 and 2 being the second most unintelligible.

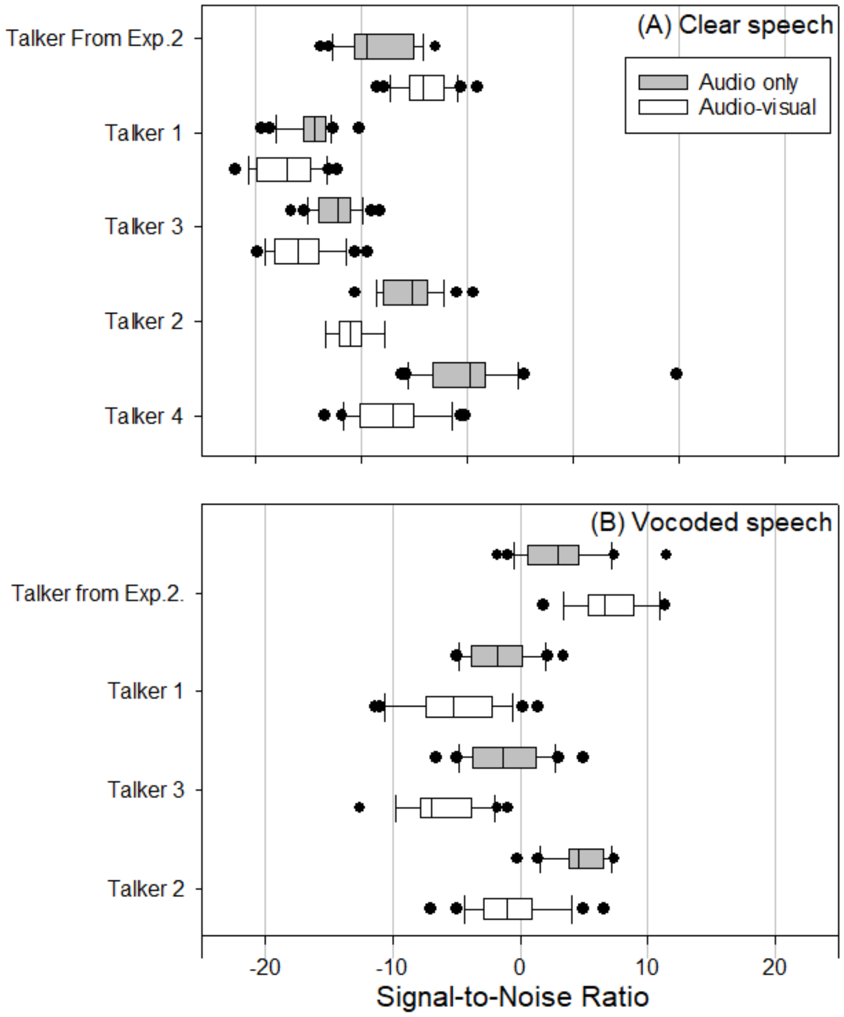


Figure 8.1 Audio and audio-visual performance levels for the talker from Experiment 2 in the sixteen talker background noise and the results of the four talkers from Experiment 3 in clear speech (A) and the three talkers in vocoded speech (B).

Figure 8.2 shows the amount of visual speech benefit each talker provided, arranged according to their audio intelligibility in clear speech. This shows that although the target talker from Experiments 1 and 2 was the second most unintelligible, the amount of visual speech benefit provided was not monotonically related to the audio intelligibility as was found for the other target talkers in Experiment 3 in clear speech. For the other target talkers in Experiment 3, the amount for visual speech benefit increased as the audio intelligibility deceased. This suggests this may be a feature of this particular talker, the visual speech information provided by this talker is less than provided by other talkers. It could also be argued this may be a result of the test conditions used in Experiment 2 when compared to Experiment 3. Greater experience was gained with the target talker by participants in Experiment 2, where thirteen conditions (260 sentences) with the same talker were presented to each participant. This is in contrast to Experiment 3, where each target talker was seen for four conditions (80 sentences). This could suggest that greater familiarity with the target reduced the amount of visual speech benefit gained. This could be because greater experience with the audio only speech for this talker increased familiarity with the talker, leading to a reduced opportunity for gain once the visual information was added to the auditory stream; a ceiling effect may be reached. The amount of the visual speech benefit gained may vary as a result of the amount of experience gained with a target talker. Future research could examine whether the amount of visual speech benefit provided by target talkers changes as a result of experience with each talker.

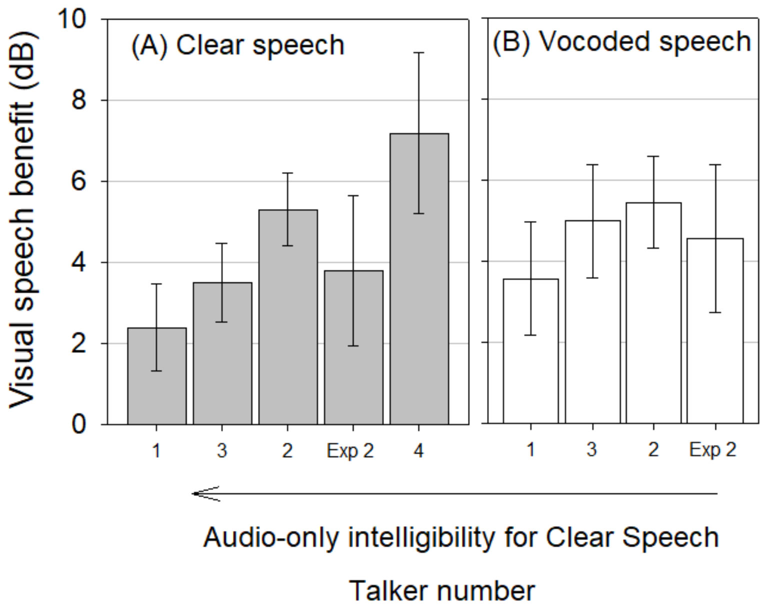


Figure 8.2 The amount of visual speech benefit for the target talker from Experiments 1 and 2 compared to the four target talkers in Experiment 3. Results are shown for clear speech in a sixteen talker background noise. Talkers are ordered according to the audio only intelligibility in clear speech.

8.2 Comparison of Results for Experiment 4 and 5

Experiment 4 assessed the predictors that may explain the significant variation in the amount of visual speech benefit received by participants. For clear speech, there were three significant predictors of the amount of visual speech benefit gained.

These were; general speech perception ability, ability to detect audio and visual synchrony and tendency towards autistic traits. Experiment 5 measured these same variables but no significant predictors of the amount of visual speech benefit received were found. This is in contrast to the results of Experiment 4 where a significant model was observed explaining 60% of the variance in the amount of visual speech benefit received.

However, the visual speech benefit data for Experiment 5 (M=4.15dB, SD=2.7) showed a trend similar to that of Experiment 4 (M=3.29dB, SD=3). Figure 8.3 shows the amount of visual speech benefit received in Experiments 4 and 5 in ascending order of the amount of benefit. Figure 8.4 shows the distribution of the mean amount of visual speech benefit in Experiments 4 and 5. These show the similarity between the two data sets.

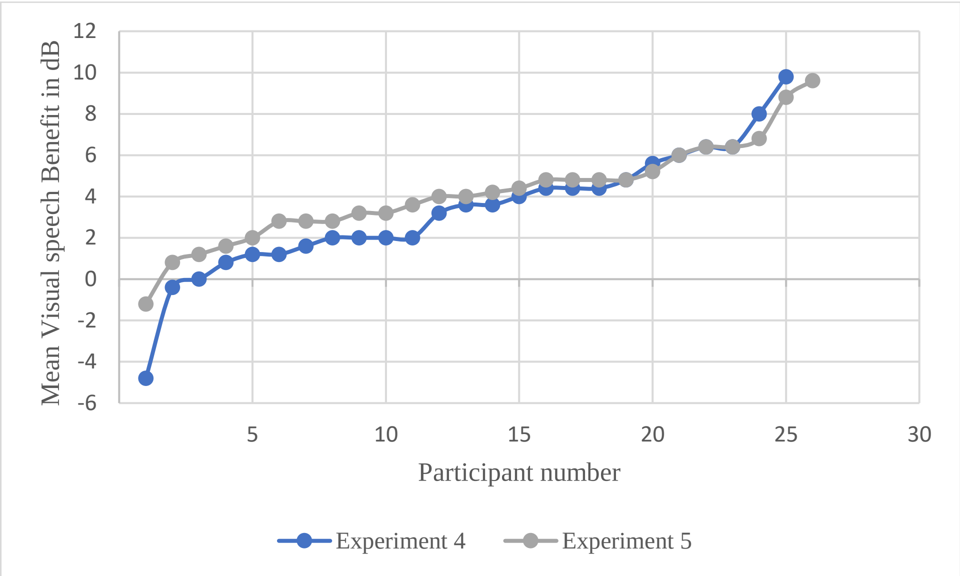


Figure 8.3 The amount of visual speech benefit for clear speech in Experiment 4 and Experiment 5 in ascending order.

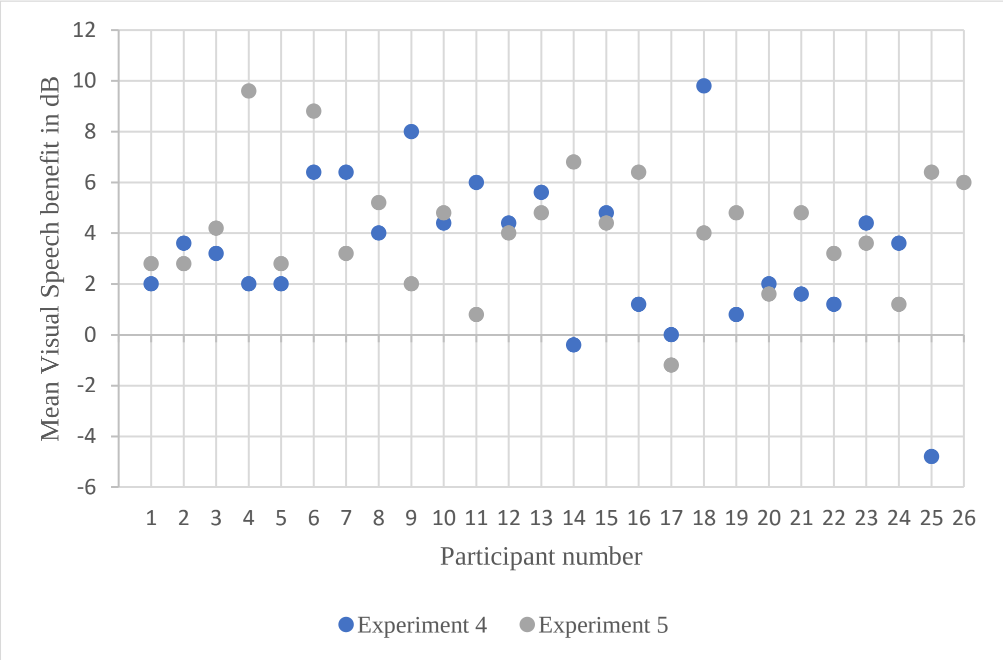


Figure 8.4 The distribution of the mean amount of visual speech benefit for each participant in Experiment 4 and for the different participants in Experiment 5 for clear speech.

8.2.1 Comparison of Results across Experiments 4 and 5 for Clear Speech

In order to further assess the results of Experiment 4 and 5, the regression analysis was re-run combining the data from Experiment 4 and 5 for clear speech. However, when comparing the full data sets, one significant difference between them was found.

The results of the audio and visual synchrony task in Experiment 5 (M=61% correct, SD=10.1) were lower than in Experiment 4 (M=77% correct, SD=12.3). The audio and visual synchrony task in Experiment 5 was run as two separate blocks compared to one continuous block used in Experiment 4. This was because of technical difficulties with the E-Prime software. Consequently, a combined regression analysis including this measure was not possible and was therefore carried out without the audio and visual synchrony results. The remaining data for Experiments 4 and 5 were combined and are discussed in the next section.

The Regression Analysis for Clear Speech for Experiments 4 and 5

The average amount of visual speech benefit for clear speech over the two experiments was 3.7dB (SD=2.7). A regression analysis was run on the combined results of Experiment 4 and 5 (N=49). The regression model was significant and showed that the predictor variables explained 26% of the variance in the amount of visual speech benefit received (adj. R^2=.26, $F(2,48)$=9.9, p<.001). Better general speech perception ability predicted a higher level of visual speech benefit (β=-.45, p=.001) and a higher score on the autism spectrum quotient (AQ) predicted a lower level of visual speech benefit (β=-.3, p=.02) (Table 8.1).

Table 8.1 The regression analysis for clear speech for Experiments 4 and 5

Variable	B	SE(B)	β	t	p
AQ	-.11	0.05	-.3	-2.44	.02
General Speech Perception	-.74	.2	-.45	-3.69	.001

These combined results show that although the results of Experiment 5 did not replicate the regression results of Experiment 4, the overall trends in the data were similar. This is shown by the similarity in the distribution of the amount of visual speech benefit (Figures 8.1 and 8.2) and by the results of the regression analysis across both Experiment 4 and Experiment 5.

Further regression analysis was then run without the general speech perception ability measure. As outlined in section 6.2.1, this measure was added to the analysis to take account of the possible confounds of participants having greater scope for improvement in task performance if they were poorer at the task initially. As also discussed in section 6.4.4, modelling of the data suggests that if there is unequal variance between the AO and AV performance levels then some correlation with visual speech benefit is to be expected, given that visual speech benefit is calculated by subtracting the audio visual performance level from the audio-only performance level.

The regression analysis was therefore run using the combined data of Experiment 4 and 5 to predict if the amount of visual speech benefit was predicted by the tendency towards greater autistic traits, as measured by the autism spectrum quotient (AQ). The regression was significant, (adj. R^2=.073, $F(1,49)$=4.94, p=.03) with the score on the AQ significantly predicting the amount of visual speech benefit gained in clear speech (β=-.3, p=.02). This suggests the tendency towards autistic traits predicts the amount of visual speech benefit gained, regardless of overall performance on the task. It is not known whether this significance would remain should the target talker or the background noise change. Future research could examine whether the tendency towards greater autistic traits reduces the amount of visual speech benefit gained for all target talkers and in all types of background noise.

8.3 Chapter Summary

- The target talker used in Experiment 1 and 2 was the second most unintelligible talker used in this project
- The amount of visual speech benefit provided by the target talker from Experiments 1 and 2 did not relate to the audio intelligibility in clear speech, as found for the four other target talkers used in Experiment 4. The talker did not provide as much visual speech benefit as would be expected from the lower level of audio intelligibility of this talker
- The regression model for clear speech from Experiment 5 was not significant, however, further examination of the data shows that the trends in the data were similar to that of Experiment 4, where a significant regression model was found
- When combined, the data from Experiments 4 and 5 for clear speech had a significant regression model, with general speech perception ability and lower tendency towards autistic traits predicting the greater amounts of visual speech benefit
- Further analysis for clear speech, without the general speech perception ability measure, showed tendency towards autistic traits remained a significant predictor of the amount of visual speech benefit received.

Chapter 9 – General Summary and Future Research

Chapter 9 summarises the main findings and implications of the five experiments outlined in this project. Ideas for future research are then discussed.

9.1 The Aims of this Project

The aims of this project were to assess the factors that change the amount of visual speech benefit individuals are able to gain from visual information when understanding speech in background noise. This was examined for clear speech and for vocoded speech, which simulates the type of speech experienced by cochlear implants users. Changes to the speech source were assessed, as well as differences between the individuals perceiving the speech. Overall therefore, external and internal changes were examined. External changes to the speech examined variation in the amount of visual speech benefit that was gained if the number of talkers in the background noise varied and if the target talker varied in intelligibility. Internal differences were examined with reference to individual differences between participants that changed the amount of visual speech benefit gained.

9.2 General Summary of Experimental Results

9.2.1 Summary of Experiments 1, 2 and 3

The main results of Experiment 1, 2 and 3, which examined changes to the speech stimuli, are summarised in this section.

Experiment 1 examined the variability in the amount of visual speech benefit gained if different types of background noise were used in the test stimuli; one talker, two talkers and sixteen talkers, for clear speech and for vocoded speech. A significant variation in performance was found between individuals, but no difference in the amount of visual speech benefit received was found between the background noise types or between clear and vocoded speech. Performance levels in Experiment 1 were not as high as expected, based on previous research, suggesting participants were experiencing difficulty identifying the target speech from the background noise. In order to address this difficulty, Experiment 2 repeated the methodology of Experiment 1 and included an inter aural delay to allow participants to identify the target voice more easily and achieve better performance levels.

The results of Experiment 2 showed that audio only and audio-visual performance declined with the introduction of more background talkers in clear speech, however for vocoded speech, the increase in the number of background talkers did not change performance levels. This could suggest masking effects are operating more effectively in vocoded speech, resulting in little opportunities for glimpsing, and resulting in floor effects with very few background talkers. In addition, the amount of visual speech benefit gained in Experiment 2 did not vary significantly as the type of background noise changed, for both clear speech and vocoded speech. There was no difference between the amount of visual speech benefit gained for clear speech and vocoded speech. It was expected that greater levels of visual speech benefit would be found for vocoded speech in line with the Principle of Inverse Effectiveness (Pof IE). The PofIE suggests that multi-sensory integration is enhanced as unimodal performance declines. The results of Experiment 2 suggest that the PofIE may not apply equally to all visual speech perception tasks.

Experiment 3 examined changes to the amount of visual speech benefit gained if the target talker used in the stimuli varied. It is known that talkers vary in audio intelligibility, but it is not known whether this variability changed the amount of visual speech benefit received by participants. In Experiment 3, four target talkers were assessed for audio only, audio-visual, visual only performance levels and the amount of visual speech benefit they provided. Significant variation in intelligibility between talkers was found in audio only and audio-visual conditions. In clear and vocoded speech, the amount of visual speech benefit generally increased as the audio intelligibility of the target talker decreased in clear speech, in line with the Principle of Inverse Effectiveness. As the unimodal performance level declined, with the decrease in audio intelligibly of the target talker, multisensory integration was enhanced, and more visual speech benefit gained. However, this did not hold true for all talkers in vocoded speech. In addition, there was a greater amount of visual speech benefit provided in vocoded speech than in clear speech overall, but again, this did not hold for all target talkers. This suggests the Principle of Inverse Effectiveness, may not apply monotonically to visual speech perception tasks and may vary depending on the test stimuli used in the research.

9.2.2 Summary of Experiments 4 and 5

The main results of Experiment 4 and 5, assessing individual differences between participants are summarised in this section.

There was significant variation in the amount of visual speech benefit received between participants. Individual differences that may predict the variation in the amount of visual speech benefit gained were measured in Experiment 4. The individual difference measures were; general speech perception ability, sensitivity to temporal fine structure information, aspects of attention, working memory capacity, ability to detect audio and visual synchrony, and tendency towards general autistic traits. This was measured for both clear and vocoded speech. For clear speech, three of the predictors explained 60% of the variance in the amount of visual speech benefit received. The significant predictors were; general speech perception performance, ability to detect audio and visual synchrony and tendency towards autistic spectrum traits. For vocoded speech, there were no significant predictors of the amount of visual speech benefit gained. Additionally, there was no difference in the amount of visual speech benefit gained between clear and vocoded speech. Performance in vocoded speech tasks was more varied than in clear speech tasks, in line with the previous experiments in this project. Moreover, there was a greater change in performance over time in the vocoded speech tasks than the clear speech tasks. Taken together, this suggests less experience with vocoded speech may be contributing to variation in performance levels. Greater experience with clear speech may be leading to more stable performance levels. Greater experience with vocoded speech stimuli was added to the next study, Experiment 5.

Experiment 5 repeated the speech perception tasks and measured the same individual difference measures from Experiment 4, with the addition of measurement of gaze direction and duration measured by eye tracking. Additionally, a familiarisation task was included to stabilise performance in the vocoded speech tasks. The results of Experiment 5 showed that for clear speech, two of the predictors explained 50% of the variance in the amount of visual speech benefit gained, general speech perception ability and time spent looking at the mouth area. The familiarisation task stabilised performance over time but overall performance levels did not improve in the vocoded speech conditions. For vocoded speech, there were no significant predictors of the amount of visual speech benefit gained. There was no difference in the amount of visual speech benefit gained between clear and vocoded speech.

9.3 The Implications of this Research

9.3.1 The Benefit to Speech Understanding from Visual Information

Generally, speech understanding was improved throughout this project by the inclusion of visual information to the auditory stream of information. However, the first experiment in this project found that this is not the case for all speech perception tasks.

In Experiment 1, significantly more speech could be understood if audio-visual information was presented rather than audio information alone. There was no significant interaction effect between modality, speech type and background noise type. The addition of visual information did not significantly improve speech perception. Additionally, changes to the number of talkers in the background did not significantly change the amount of visual speech benefit gained by participants. This shows that the additional visual information provided in the audio-visual conditions was not enough to improve performance levels, despite the benefits that visual information is known to provide in speech understanding (e.g. Middelweerd & Plomp, 1987. In Experiment 1, the additional visual information was not enough to overcome the challenge of identifying the target voice from the background noise and therefore participants were unable to report the target voice, resulting in floor effects.

This suggests that although visual information is generally agreed to aid speech perception, this may not always provide enough benefit to overcome particularly unintelligible speech. As discussed in section 8.1, the talker used in Experiment 1 was found to be relatively unintelligible compared to other talkers used in this project. To overcome the particular challenges of understanding the talker used in Experiment 1, Experiment 2 incorporated spatial cues in the form of an interaural delay. This was designed to enable participants to more easily identify the target talker from the background noise. The inclusion of the interaural delay in Experiment 2 was successful in raising the threshold at which the visual information was of benefit to participants, more speech could be understood if audio-visual information was presented rather than auditory information alone in all conditions.

It is also worth noting that performance in audio-visual conditions for vocoded speech in Experiment 2 was relatively unaffected by the inclusion of more background talkers. This again suggests the information provided by the visual input was not enough to support greater speech understanding in vocoded speech conditions. A limit was reached with a small number of background talkers. Overall, this suggests that visual information is only of benefit to speech perception once a certain audio threshold of understanding is achieved. It is therefore important to consider the particular challenges of the speech task participants are being asked to undertake in research with regard to the relative intelligibility of the target talker and the challenges the background noise may add to the task. In real-world speech understanding it may be worth highlighting to listeners that the addition of visual information may not always be enough to overcome particularly degraded speech. This may apply equally to normal hearing and hearing-impaired listeners.

Maximising Visual Speech Benefit

In the first 3 experiments in this project, large variability in performance levels was found between participants. Experiments 4 and 5 examined this variability with regard to individual differences between participants. For clear speech in Experiment 4, a ‘super visual-speech perceiver’ was identified. Expertise in general speech performance and in detecting AV synchrony, and a lower score on the AQ, lead to better overall ability to obtain visual speech benefit. The results of Experiment 5 found a similar pattern of results to those found in Experiment 4.

Arguably, the significant predictors of the amount of visual speech benefit received are innate; general speech perception ability, ability to detect AV synchrony and score on the AQ. This would suggest consideration of these abilities and traits should be made when conducting research that assesses visual speech perception. This becomes more critical where smaller sample sizes are used in research. This project assessed speech perception in background noise. It is not known whether these predictors remain significant if the background is removed or is of a different type.

Experiment 5 found greater gaze time on the mouth area significantly increased the amount of visual speech benefit gained in clear speech. Gaze patterns did not change as a result of changes to the SNR and as a function of accuracy, and therefore it is possible that gaze patterns are driven by individual differences between participants. This would suggest instructing participants and possibly listeners in general to increase gaze time towards the mouth would be beneficial. The eye tracking evidence from Experiment 5 suggest that this strategy is not being used by all participants. This may also be a useful strategy for hearing impaired populations to employ.

It may also be possible to maximise the amount of visual speech benefit gained by participants through training. Experiment 5 showed that a limited training task was able to stabilise performance for vocoded speech. It may be possible to improve performance levels with further training in order that comparable performance levels can be achieved in clear and vocoded speech tasks. The gaze strategy that improved speech perception in clear speech may then be more beneficial in vocoded speech tasks.

9.3.2 The Selection of Research Stimuli

This research used IEEE sentences throughout the project. These sentences are less predictable than other speech materials and may be more challenging for listeners than single word lists or less complex sentences used in other research (e.g. Altieri and Wenger, 2013; Ross et al., 2006; Sumby & Pollack, 1954). There are also known increased masking effects from the type of background noise used in this project (speech noise) than other background noise types, such as pink or white noise (e.g. Brungart, 2001). Therefore, the choice of stimuli may change the amount of visual speech benefit that listeners are able to gain.

Additionally, the level of benefit was measured using an adaptive track procedure in this research and not using a fixed performance level as used by other researchers (e.g. Altieri and Wenger, 2013; Ross et al., 2006). Moreover, Experiment 3 highlighted the significant differences in intelligibility between talkers and the changes to the amount of visual speech benefit gained as a result of this.

It is therefore important consider the stimuli selected for use in any research and the methodology used to assess performance levels, in order that accurate performance levels are measured. Additionally, it is suggested the same talker be used throughout a research study to ensure accurate measurement of performance levels is achieved between conditions and participants. The relative intelligibility of the target talker should also be assessed to ensure consistency across research studies. As highlighted in Experiment 3, talkers varied in their levels of intelligibility by 15db in clear speech and 17dB in vocoded speech in audio only conditions. The amount of visual speech benefit the talkers provided also varied.

As outlined in section 5.5.1, variability in testing stimuli would also be particularly problematic for testing cochlear implant (CI) users. If it is assumed the vocoding of stimuli in this experiment is an accurate proximity of the type of audio information that CI users receive, variation in target talkers would significantly change hearing assessment results. If testing stimuli differ, comparison across time for patients and between patient groups becomes problematic. Given the large variation in patients' outcomes (Pisoni et al., 2016) and the importance of CI implants for patients' long-term health outcomes (Vermeire et al., 2005) it is important to ensure accurate and consistent measurement of performance. This is important for managing CI users' expectations and to assess the relative benefits of the implants.

9.3.3 The Principle of Inverse Effectiveness

It was hypothesised in this project that the Principle of Inverse Effectiveness (PofIE) may apply to the integration of audio-visual stimuli and the amount of visual speech benefit received. It is argued that as speech intelligibility is reduced due to the addition of background noise, the contribution of visual information uniformly increases (Sumby & Pollack, 1954). Similarly, where the audio stream of information is most degraded, additional information provided by the visual information stream may be at its most beneficial (Meredith and Stein, 1986). Therefore, where speech is degraded by vocoding, the contribution of the visual information would be at its greatest. However, the results of the experiments in this project were not entirely consistent with this principle.

The results of Experiment 2 showed that for the target talker used in this experiment, there was no difference in the amount of visual speech benefit gained in clear and vocoded speech. This was also found for the talker used in Experiments 4 and 5, there was no difference in the amount of visual speech benefit gained between clear and vocoded speech in these two experiments. If the PofIE were to apply, it would be expected that greater gain would be found in vocoded speech conditions. Similarly, more visual speech benefit would be expected as the number of background talkers increased, and the speech became more degraded. Again, this was not found in Experiment 2.

In contrast, participants received more visual speech benefit overall for vocoded speech conditions than in clear speech in Experiment 3. However, this did not apply for all target talkers used in this experiment. There was no significant difference found between the amount of visual speech benefit in clear and vocoded speech for Talker 2. Similarly, the amount of visual speech benefit found in Experiment 3 increased as the audio intelligibility of the target talker decreased in clear speech. However, for vocoded speech the amount of visual speech benefit gained did not increase uniformly with the decrease in intelligibility of the target talker.

Therefore, overall, this would suggest the PofIE does not apply as monotonically to speech perception as proposed by Sumby and Pollack (1954) but has a more complex relationship with the integration of auditory and visual information. Variation in the target talker or background noise, changes the nature of the speech integration. This may indicate there is a limit to the benefit that the PofIE confers. This in line with previous research (Altieri and Wenger, 2013; Ross et al. 2006), where maximal limits to the amount of visual speech benefit that could be gained were found.

9.4 Future Research

The results of the first experiment in this project showed that for the particular stimuli used, it was particularly challenging to identify the target voice from the background noise. An interaural delay was added to the stimuli for the following experiment, Experiment 2, and for the remaining experiments in the project, to enable listeners to more easily identify the target voice. The target stimuli were IEEE sentences, which contain limited contextual information, and the background noise was other talkers. Background noise consisting of other talkers is known to have increased informational masking effects than other types of background noise, such as pink or white noise. Using this type of stimuli, it was found that for clear speech, audio and audio-visual performance levels decreased as the number of background talkers increased. However, for vocoded speech, there was no change in performance levels as the number of background talkers increased. This suggests floor levels of performance had been reached with one background talker for vocoded speech. Repeating this experiment with different target talkers and with different sentence stimuli would establish whether this floor effect was a feature of the target talker used in Experiment 2 or a feature of the test stimuli. Sentence stimuli with more contextual information may lead to improvements in performance, as may changing the background noise type. Moreover, as found in Experiment 3, different talkers have varied levels of audio intelligibility, resulting in difference in performance levels. Repeating this experiment with varied target talkers with different test stimuli would establish possible limits to performance levels in vocoded speech. The Experiments in this project also measured performance levels using an adaptive track procedure and not fixed performance levels. Further research could examine performance levels using a fixed SNR to establish if different talkers and different stimuli have an optimum SNR at which the maximum amount of visual speech benefit is gained. This may be dependent on the test stimuli used and the audio intelligibly of the target talker.

Similarly to performance in the audio and audio-visual tasks, the amount of visual speech benefit gained by individuals was measured in this project using a particular set of stimuli and background noise. Experiment 3 showed that for some talkers, greater levels of visual speech benefit were found in vocoded speech than in clear speech. This was not found for other talkers. It was also observed that the amount of visual speech benefit gained in clear speech increased as the audio intelligibility decreased in clear speech. This was not found for vocoded speech. However, this limit could be a result of the test stimuli used. Using more contextual sentences and changing the type of background noise may result in greater amounts of visual speech benefit in vocoded speech over clear speech. Future studies could repeat this experiment with differing test stimuli to establish where the greatest amount of visual speech benefit is gained and for which talkers.

Experiment 4 and 5 in this project examined individual differences between participants that would predict the level of visual speech benefit gained. This was examined using a single target talker and the same test stimuli; IEEE sentences and background noise of sixteen talkers. It is unclear whether the significant predictors of the amount of visual speech benefit gained would remain significant using different test stimuli. A different target talker, with greater or less intelligibility, sentences with greater context and differing types of background noise may change this model. Further testing to check the robustness of the significant model could be carried out.

It was also argued that prior exposure to a particular talker may lead to increased visual speech benefit as familiarity with that talker increases. This may vary as a result of the intelligibility of the target talker. Talkers who have particularly low levels of intelligibility may provide a greater increase in performance levels as participants become more familiar with the individual characteristics of that talker. Further research is needed to establish if experience with a target talker changes the amount of visual speech benefit gained.

Experience with vocoded speech stimuli was given to participants in Experiment 5 in this project. However, there was not enough experience gained by some participants to enable them to achieve a more similar performance in the clear speech and vocoded speech conditions as expected. It could be argued that substantially more training may change performance levels in vocoded speech to be more comparable to those achieved in clear speech. However, it may not be possible through training to achieve the same level of expertise with vocoded speech as clear speech, given that participants would have spent a life time gaining experience with clear speech. In order to address this, the experiment could be repeated with experienced cochlear implant (CI) users who are familiar with vocoded speech. Performance levels could then be compared between CI users and normal hearing listeners with training, to establish if expertise with vocoded speech can be achieved. It would then be possible to compare performance levels for participants who have acquired vocoded speech expertise, with their performance in clear speech. If this resulted in similar performance levels, differences in the amount of visual speech benefit gained could be attributed to the speech type and not affected by the level of experience with the test stimuli.

The training of participants with vocoded speech stimuli would also be useful to carry out so that Experiment 4 and 5 could be retested on this group. The significant predictors of the amount of visual speech benefit received in clear speech were not found in the vocoded speech conditions. This could be a result of inexperience with the test stimuli and the particular challenges of the target talker and background noise. It would be useful to train and then retest participants using varied test stimuli to establish if the significant predictors of the amount of visual speech benefit gained hold regardless of the test conditions.

Finally, the results of Experiment 5 show that focus on the mouth area during speech perception results in greater gains from visual speech. It would be interesting to test this prediction using different speech stimuli, with greater contextual information, with different, and perhaps less challenging, background noise, and with different target talkers. If this strategy holds regardless of the test conditions, this would be a useful intervention strategy to relay to listeners who are experiencing difficulties perceiving speech in background noise. Focus on the mouth area results in additional gains from visual speech regardless of the context.

9.5 Conclusion

Overall it can be seen that the amount of visual speech benefit gained by participants varies as a result of the type of background noise, the speech type, the intelligibility of the target talker and as a result of individual differences between participants. Although visual speech does generally provide significant benefit to speech understanding in background noise, for both clear and vocoded speech, in particularly unintelligible speech, the visual information does not provide enough benefit to allow significant gains in speech understanding.

References

Adolphs, R. (2002). Recognizing emotion from facial expressions: psychological and neurological mechanisms. *Behavioral and cognitive neuroscience reviews*, *1*(1), 21-62.

Alais, D., & Burr, D. (2004). The ventriloquist effect results from near-optimal bimodal integration. *Current biology*, *14*(3), 257-262.

Alcántara, J. I., Weisblatt, E. J., Moore, B. C., & Bolton, P. F. (2004). Speech-in-noise perception in high-functioning individuals with autism or Asperger's syndrome. *Journal of Child Psychology and Psychiatry*, *45*(6), 1107-1114.

Alsius, A., Wayne, R. V., Paré, M., & Munhall, K. G. (2016). High visual resolution matters in audiovisual speech perception, but only for some. *Attention, Perception, & Psychophysics*, *78*(5), 1472-1487.

Altieri, N. A., Pisoni, D. B., & Townsend, J. T. (2011). Some normative data on lip-reading skills (L). *The Journal of the Acoustical Society of America*, *130*(1), 1-4.

Altieri, N., & Wenger, M. J. (2013). Neural dynamics of audiovisual speech integration under variable listening conditions: an individual participant analysis. *Frontiers in psychology*, *4*.

Anderson, C. A., Wiggins, I. M., Kitterick, P. T., & Hartley, D. E. (2017). Adaptive benefit of cross-modal plasticity following cochlear implantation in deaf adults. *Proceedings of the National Academy of Sciences*, *114*(38), 10256-10261.

Anderson, S., White-Schwoch, T., Parbery-Clark, A., & Kraus, N. (2013). A dynamic auditory-cognitive system supports speech-in-noise perception in older adults. *Hearing research*, *300*, 18-32.

Auer, E. T., & Bernstein, L. E. (2007). Enhanced visual speech perception in individuals with early-onset hearing impairment. *Journal of Speech, Language, and Hearing Research*, *50*(5), 1157-1165.

Badri, R., Siegel, J. H., & Wright, B. A. (2011). Auditory filter shapes and high-frequency hearing in adults who have impaired speech in noise performance despite clinically normal audiograms. *The Journal of the Acoustical Society of America*, *129*(2), 852-863.

Baer, T., Moore, B. C., & Gatehouse, S. (1993). Spectral contrast enhancement of speech in noise for listeners with sensorineural hearing impairment: effects on intelligibility, quality, and response times. *Journal of rehabilitation research and development, 30*, 49-49.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., ... & Treiman, R. (2007). The English lexicon project. *Behavior research methods*, *39*(3), 445-459.

Baron-Cohen, S., Wheelwright, S., & Jolliffe, A. T. (1997). Is there a" language of the eyes"? Evidence from normal adults, and adults with autism or Asperger syndrome. *Visual Cognition*, *4*(3), 311-331.

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The "Reading the Mind in the Eyes" Test revised version: a study with normal adults, and adults with Asperger syndrome or high-functioning autism. *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, *42*(2), 241-251.

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The autism-spectrum quotient (AQ): Evidence from asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *Journal of autism and developmental disorders*, *31*(1), 5-17.

Baskent, D., & Bazo, D. (2011). Audiovisual asynchrony detection and speech intelligibility in noise with moderate to severe sensorineural hearing impairment. *Ear and hearing*, *32*(5), 582-592.

Bench, J., Kowal, Å., & Bamford, J. (1979). The BKB (Bamford-Kowal-Bench) sentence lists for partially-hearing children. *British journal of audiology*, *13*(3), 108-112.

Bent, T., Buchwald, A., & Pisoni, D. B. (2009). Perceptual adaptation and intelligibility of multiple talkers for two types of degraded speech. *The Journal of the Acoustical Society of America*, *126*(5), 2660-2669.

Bernstein, J. G., & Oxenham, A. J. (2006). The relationship between frequency selectivity and pitch discrimination: Sensorineural hearing loss. *The Journal of the Acoustical Society of America*, *120*(6), 3929-3945.

Bernstein, L. E., & Liebenthal, E. (2014). Neural pathways for visual speech perception. *Frontiers in neuroscience*, *8*, 386.

Bernstein, L. E., Tucker, P. E., & Demorest, M. E. (2000). Speech perception without hearing. *Perception & Psychophysics*, *62*(2), 233-252.

Bizley, J. K., & Cohen, Y. E. (2013). The what, where and how of auditory-object perception. *Nature Reviews Neuroscience*, *14*(10), 693.

Blamey, P., Arndt, P., Bergeron, F., Bredberg, G., Brimacombe, J., Facer, G., ... & Shipp, D. (1996). Factors affecting auditory performance of postlinguistically deaf adults using cochlear implants. *Audiology and Neurotology*, *1*(5), 293-306.

Blamey, P., Artieres, F., Başkent, D., Bergeron, F., Beynon, A., Burke, E., ... & Govaerts, P. J. (2013). Factors affecting auditory performance of postlinguistically deaf adults using cochlear implants: an update with 2251 patients. *Audiology and Neurotology*, *18*(1), 36-47.

Bond, Z. S., & Moore, T. J. (1994). A note on the acoustic-phonetic characteristics of inadvertently clear speech. *Speech communication*, *14*(4), 325-337.

Boothroyd, A., Hanin, L., & Hnath-Chisolm, T. (1985) *A sentence test of speech perception: Reliability, set-equivalence, and short-term learning.* New York: City University of New York.

Bradlow, A. R., Torretta, G. M., & Pisoni, D. B. (1996). Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. *Speech communication*, *20*(3-4), 255-272.

Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *The Journal of the Acoustical Society of America*, *109*(3), 1101-1109.

Brungart, D. S., & Simpson, B. D. (2007). Cocktail party listening in a dynamic multitalker environment. *Perception & psychophysics*, *69*(1), 79-91.

Buchan, J. N., Paré, 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.

Calvert, G. A., Bullmore, E. T., Brammer, M. J., Campbell, R., Williams, S. C., McGuire, P. K., ... & David, A. S. (1997). Activation of auditory cortex during silent lipreading. *science*, *276*(5312), 593-596.

Carbonell, K. M. (2017). Reliability of individual differences in degraded speech perception. *The Journal of the Acoustical Society of America*, *142*(5), EL461-EL466.

Carhart, R., Johnson, C., & Goodman, J. (1975). Perceptual masking of spondees by combinations of talkers. *The Journal of the Acoustical Society of America*, *58*(S1), S35-S35.

Carlile, S. (1996). The physical and psychophysical basis of sound localization. In *Virtual auditory space: Generation and applications* (pp. 27-78). Berlin: Springer.

Chang, Y. P., & Fu, Q. J. (2006). Effects of talker variability on vowel recognition in cochlear implants. *Journal of Speech, Language, and Hearing Research*, *49*(6), 1331-1341.

Chaste, P., & Leboyer, M. (2012). Autism risk factors: genes, environment, and gene-environment interactions. *Dialogues in clinical neuroscience*, *14*(3), 281.

Chen, L. C., Puschmann, S., & Debener, S. (2017). Increased cross-modal functional connectivity in cochlear implant users. *Scientific reports*, *7*(1), 10043.

Chen, F. S., & Yoon, J. M. (2011). Brief report: broader autism phenotype predicts spontaneous reciprocity of direct gaze. *Journal of autism and developmental disorders*, *41*(8), 1131-1134.

Chita-Tegmark, M. (2016). Attention allocation in ASD: a review and meta-analysis of eye-tracking studies. *Review Journal of Autism and Developmental Disorders*, *3*(3), 209-223.

Conrey, B., & Gold, J. M. (2006). An ideal observer analysis of variability in visual-only speech. *Vision research, 46*(19), 3243-3258.

Conrey, B., & Pisoni, D. B. (2006). Auditory-visual speech perception and synchrony detection for speech and nonspeech signals. *The Journal of the Acoustical Society of America*, *119*(6), 4065-4073.

Cooke, M. (2006). A glimpsing model of speech perception in noise. *The Journal of the Acoustical Society of America*, *119*(3), 1562-1573.

Cox, R. M., Alexander, G. C., & Gilmore, C. (1987). Intelligibility of average talkers in typical listening environments. *The Journal of the Acoustical Society of America*, *81*(5), 1598-1608.

Cullington, H. E., & Zeng, F. G. (2008). Speech recognition with varying numbers and types of competing talkers by normal-hearing, cochlear-implant, and implant simulation subjects. *The Journal of the Acoustical Society of America*, *123*(1), 450-461.

Dai, L., Best, V., & Shinn-Cunningham, B. G. (2018). Sensorineural hearing loss degrades behavioral and physiological measures of human spatial selective auditory attention. *Proceedings of the National Academy of Sciences*, *115*(14), E3286-E3295.

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.

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.

Demorest, M. E., & Bernstein, L. E. (1992). Sources of variability in speechreading sentences: A generalizability analysis. *Journal of Speech, Language, and Hearing Research*, *35*(4), 876-891.

Dhanasingh, A., & Jolly, C. (2017). An overview of cochlear implant electrode array designs. *Hearing research*, *356*, 93-103.

Diederich, A., & Colonius, H. (2004). Bimodal and trimodal multisensory enhancement: effects of stimulus onset and intensity on reaction time. *Perception & psychophysics*, *66*(8), 1388-1404.

Dixon, N. F., & Spitz, L. (1980). The detection of auditory visual desynchrony. *Perception, 9*(6), 719-721.

Dodd, B., Plant, G., & Gregory, M. (1989). Teaching lip-reading: The efficacy of lessons on video. *British journal of audiology*, *23*(3), 229-238.

Donohue, S. E., Darling, E. F., & Mitroff, S. R. (2012). Links between multisensory processing and autism. *Experimental Brain Research*, *222*(4), 377-387.

Dorman, M. F., Loizou, P. C., & Rainey, D. (1997). Speech intelligibility as a function of the number of channels of stimulation for signal processors using sine-wave and noise-band outputs. *The Journal of the Acoustical Society of America*, *102*(4), 2403-2411.

Eaves, J. M., Summerfield, Q. A., & Kitterick, P. T. (2011). Benefit of temporal fine structure to speech perception in noise measured with controlled temporal envelopes. *The Journal of the Acoustical Society of America*, *130*(1), 501-507.

Eckert, M. A. (2011). Slowing down: age-related neurobiological predictors of processing speed. *Frontiers in neuroscience*, *5*, 25.

Erber, N. P. (1969). Interaction of audition and vision in the recognition of oral speech stimuli. *Journal of Speech, Language, and Hearing Research*, *12*(2), 423-425.

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

Evans, S., McGettigan, C., Agnew, Z. K., Rosen, S., & Scott, S. K. (2016). Getting the cocktail party started: masking effects in speech perception. *Journal of cognitive neuroscience*, *28*(3), 483-500.

Falck-Ytter, T., Nyström, P., Gredebäck, G., Gliga, T., Bölte, S., EASE team, ... & Hedenius, M. (2018). Reduced orienting to audiovisual synchrony in infancy predicts autism diagnosis at 3 years of age. *Journal of Child Psychology and Psychiatry*, *59*(8), 872-880.

Feld, J. E., & Sommers, M. S. (2009). Lipreading, processing speed, and working memory in younger and older adults. *Journal of Speech, Language, and Hearing Research*, *52*(6), 1555-1565.

Festen, J. M., & Plomp, R. (1990). Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. *The Journal of the Acoustical Society of America*, *88*(4), 1725-1736.

Fitch, W. T., & Giedd, J. (1999). Morphology and development of the human vocal tract: A study using magnetic resonance imaging. *The Journal of the Acoustical Society of America*, *106*(3), 1511-1522.

Freeth, M., Foulsham, T., & Kingstone, A. (2013). What affects social attention? Social presence, eye contact and autistic traits. *PloS one, 8*(1), e53286.

Freyman, R. L., Balakrishnan, U., & Helfer, K. S. (2004). Effect of number of masking talkers and auditory priming on informational masking in speech recognition. *The Journal of the Acoustical Society of America*, *115*(5), 2246-2256.

Friesen, L. M., Shannon, R. V., Baskent, D., & Wang, X. (2001). Speech recognition in noise as a function of the number of spectral channels: Comparison of acoustic hearing and cochlear implants. *The Journal of the Acoustical Society of America*, *110*(2), 1150-1163.

Fu, Q. J., Chinchilla, S., & Galvin, J. J. (2004). The role of spectral and temporal cues in voice gender discrimination by normal-hearing listeners and cochlear implant users. *Journal of the Association for Research in Otolaryngology*, *5*(3), 253-260.

Fu, Q. J., Nogaki, G., & Galvin, J. J. (2005). Auditory training with spectrally shifted speech: implications for cochlear implant patient auditory rehabilitation. *Journal of the Association for Research in Otolaryngology*, *6*(2), 180-189.

Fu, Q. J., & Shannon, R. V. (1999). Recognition of spectrally degraded and frequency-shifted vowels in acoustic and electric hearing. *The Journal of the Acoustical Society of America*, *105*(3), 1889-1900.

Fu, Q. J., Shannon, R. V., & Wang, X. (1998). Effects of noise and spectral resolution on vowel and consonant recognition: Acoustic and electric hearing. *The Journal of the Acoustical Society of America*, *104*(6), 3586-3596.

Füllgrabe, C., & Rosen, S. (2016). Investigating the role of working memory in speech-in-noise identification for listeners with normal hearing. In *Physiology, Psychoacoustics and Cognition in Normal and Impaired Hearing* (pp. 29-36). Cham: Springer.

Gagne, J. P., Masterson, V., Munhall, K. G., Bilida, N., & Querengesser, C. (1994). Across talker variability in auditory, visual, and audiovisual speech intelligibility for conversational and clear speech. *Journal-Academy of Rehabilitative Audiology*, *27*, 135-158.

Géléoc, G. S., & Holt, J. R. (2014). Sound strategies for hearing restoration. *Science*, *344*(6184), 1241062.

Gibney, K. D., Aligbe, E., Eggleston, B. A., Nunes, S. R., Kerkhoff, W. G., Dean, C. L., & Kwakye, L. D. (2017). Visual distractors disrupt audiovisual integration regardless of stimulus complexity. *Frontiers in integrative neuroscience*, *11*, 1.

Glasberg, B. R., & Moore, B. C. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing research*, *47*(1-2), 103-138.

Goh, W. D., Pisoni, D. B., Kirk, K. I., & Remez, R. E. (2001). Audio-visual perception of sinewave speech in an adult cochlear implant user: a case study. *Ear and hearing*, *22*(5), 412.

Golumbic, E. M. Z., Ding, N., Bickel, S., Lakatos, P., Schevon, C. A., McKhann, G. M., ... & Poeppel, D. (2013). Mechanisms underlying selective neuronal tracking of attended speech at a “cocktail party”. *Neuron*, *77*(5), 980-991.

Gonzalez, J., & Oliver, J. C. (2005). Gender and speaker identification as a function of the number of channels in spectrally reduced speech. *The Journal of the Acoustical Society of America*, *118*(1), 461-470.

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

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. *The Journal of the Acoustical Society of America*, *103*(5), 2677-2690.

Grant, K. W., Wassenhove, V. V., & Poeppel, D. (2003). Discrimination of auditory-visual synchrony. In *AVSP 2003-International Conference on Audio-Visual Speech Processing*.

Green, K. M., Bhatt, Y. M., Mawman, D. J., O'Driscoll, M. P., Saeed, S. R., Ramsden, R. T., & Green, M. W. (2007). Predictors of audiological outcome following cochlear implantation in adults. *Cochlear implants international*, *8*(1), 1-11.

Green, K. P., & Miller, J. L. (1985). On the role of visual rate information in phonetic perception. *Perception & psychophysics*, *38*(3), 269-276.

Green, K. P., Kuhl, P. K., Meltzoff, A. N., & Stevens, E. B. (1991). Integrating speech information across talkers, gender, and sensory modality: Female faces and male voices in the McGurk effect. *Perception & psychophysics*, *50*(6), 524-536.

Green, T., Katiri, S., Faulkner, A., & Rosen, S. (2007). Talker intelligibility differences in cochlear implant listeners. *The Journal of the Acoustical Society of America*, *121*(6), EL223-EL229.

Grossman, R. B., Steinhart, E., Mitchell, T., & McIlvane, W. (2015). “Look who's talking!” Gaze Patterns for Implicit and Explicit Audio-Visual Speech Synchrony Detection in Children With High-Functioning Autism. *Autism Research*, *8*(3), 307-316.

Halley, J. M. (1996). Ecology, evolution and 1f-noise. *Trends in ecology & evolution*, *11*(1), 33-37.

Harris, M. S., Kronenberger, W. G., Gao, S., Hoen, H. M., Miyamoto, R. T., & Pisoni, D. B. (2013). Verbal short-term memory development and spoken language outcomes in deaf children with cochlear implants. *Ear and hearing*, *34*(2), 179.

Hazan, V., & Markham, D. (2004). Acoustic-phonetic correlates of talker intelligibility for adults and children. *The Journal of the Acoustical Society of America*, *116*(5), 3108-3118.

Heinrich, A., Henshaw, H., & Ferguson, M. A. (2015). The relationship of speech intelligibility with hearing sensitivity, cognition, and perceived hearing difficulties varies for different speech perception tests. *Frontiers in psychology*, *6*, 782.

Hervais-Adelman, A. G., Davis, M. H., Johnsrude, I. S., Taylor, K. J., & Carlyon, R. P. (2011). Generalization of perceptual learning of vocoded speech. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(1), 283.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, *8*(5), 393.

Hinchcliffe, R. (1992). King-Kopetzky syndrome: An auditory stress disorder. *Journal of Audiological Medicine*, *1*(2), 89-98.

Holden, L. K., Finley, C. C., Firszt, J. B., Holden, T. A., Brenner, C., Potts, L. G., ... & Skinner, M. W. (2013). Factors affecting open-set word recognition in adults with cochlear implants. *Ear and hearing*, *34*(3), 342.

Hopkins, K., & Moore, B. C. (2010). Development of a fast method for measuring sensitivity to temporal fine structure information at low frequencies. *International Journal of Audiology*, *49*(12), 940-946.

Hopkins, K., Moore, B. C., & Stone, M. A. (2008). Effects of moderate cochlear hearing loss on the ability to benefit from temporal fine structure information in speech. *The Journal of the Acoustical Society of America*, *123*(2), 1140-1153.

Humes, L. E., Kidd, G. R., & Fogerty, D. (2017). Exploring use of the coordinate response measure in a multitalker babble paradigm. *Journal of Speech, Language, and Hearing Research*, *60*(3), 741-754.

Irwin, J. R., & Brancazio, L. (2014). Seeing to hear? Patterns of gaze to speaking faces in children with autism spectrum disorders. *Frontiers in psychology*, *5*, 397.

Irwin, J., Avery, T., Turcios, J., Brancazio, L., Cook, B., & Landi, N. (2017). Electrophysiological Indices of Audiovisual Speech Perception in the Broader Autism Phenotype. *Brain Sciences*, *7*(6).

Jacoby, L. L., Allan, L. G., Collins, J. C., & Larwill, L. K. (1988). Memory influences subjective experience: Noise judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(2), 240.

Johnsrude, I. S., Mackey, A., Hakyemez, H., Alexander, E., Trang, H. P., & Carlyon, R. P. (2013). Swinging at a cocktail party: Voice familiarity aids speech perception in the presence of a competing voice. *Psychological Science*, *24*(10), 1995-2004.

Josupeit, A., & Hohmann, V. (2017). Modeling speech localization, talker identification, and word recognition in a multi-talker setting. *The Journal of the Acoustical Society of America*, *142*(1), 35-54.

Kaiser, A. R., Kirk, K. I., Lachs, L., & Pisoni, D. B. (2003). Talker and lexical effects on audiovisual word recognition by adults with cochlear implants. *Journal of Speech, Language, and Hearing Research, 46*, 390–404.

Ketten, D. R., Skinner, M. W., Wang, G. E., Vannier, M. W., Gates, G. A., & Gail Neely, J. (1998). *In vivo measures of cochlear length and insertion depth of nucleus cochlear implant electrode arrays*. Missouri: Annals Publishing Company.

Kezuka, E., Amano, S., & Reddy, V. (2017). Developmental changes in locating voice and sound in space. *Frontiers in psychology*, *8*, 1574.

Kidd Jr, G., Mason, C. R., Deliwala, P. S., Woods, W. S., & Colburn, H. S. (1994). Reducing informational masking by sound segregation. *The Journal of the Acoustical Society of America*, *95*(6), 3475-3480.

Kilman, L., Zekveld, A., Hällgren, M., & Rönnberg, J. (2014). The influence of non-native language proficiency on speech perception performance. *Frontiers in psychology*, *5*, 651.

King, K., & Stephens, D. (1992). Auditory and psychological factors in ‘auditory disability with normal hearing’. *Scandinavian audiology*, *21*(2), 109-114.

Klin, A., Jones, W., Schultz, R., Volkmar, F., & Cohen, D. (2002). Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism. *Archives of general psychiatry*, *59*(9), 809-816.

Kong, Y. Y., Donaldson, G., & Somarowthu, A. (2015). Effects of contextual cues on speech recognition in simulated electric-acoustic stimulation. *The Journal of the Acoustical Society of America*, *137*(5), 2846-2857.

Kricos, P. B., & Lesner, S. A. (1982). Differences in visual intelligibility across talkers. *The Volta Review*.

Lachs, L., & Hernandez, L. R. (1998). Update: the Hoosier audiovisual multitalker database. *Research on Spoken Language Processing Progress Report No*, *22*, 377-388.

Lai, M. C., Lombardo, M. V., Chakrabarti, B., & Baron-Cohen, S. (2013). Subgrouping the Autism “Spectrum": Reflections on DSM-5. *PLoS biology*, *11*(4), e1001544.

Lander, K., & Davies, R. (2008). Does face familiarity influence speechreadability? *The Quarterly Journal Of Experimental Psychology, 61*(7), 961-967.

Lazard, D. S., & Giraud, A. L. (2017). Faster phonological processing and right occipito-temporal coupling in deaf adults signal poor cochlear implant outcome. *Nature Communications*, *8,* 14872.

Lazard, D. S., Vincent, C., Venail, F., Van de Heyning, P., Truy, E., Sterkers, O., ... & Mawman, D. (2012). Pre-, per-and postoperative factors affecting performance of postlinguistically deaf adults using cochlear implants: a new conceptual model over time. *PloS one*, *7*(11), e48739.

Lenarz, M., Sönmez, H., Joseph, G., Büchner, A., & Lenarz, T. (2012). Long-term performance of cochlear implants in postlingually deafened adults. *Otolaryngology--Head and Neck Surgery*, *147*(1), 112-118.

Lesner, S. A. (1988). The talker. *The Volta Review*.

Leung, J., Wang, N. Y., Yeagle, J. D., Chinnici, J., Bowditch, S., Francis, H. W., & Niparko, J. K. (2005). Predictive models for cochlear implantation in elderly candidates. *Archives of Otolaryngology–Head & Neck Surgery*, *131*(12), 1049-1054.

Liberman, A. M., Cooper, F. S., Shankweiler, D. P., & Studdert-Kennedy, M. (1967). Perception of the speech code. *Psychological review*, *74*(6), 431.

Lidestam, B., Lyxell, B., & Andersson, G. (1999). Speech-reading: Cognitive predictors and displayed emotion. *Scandinavian Audiology*, *28*(4), 211-217.

Lin, F. R., Yaffe, K., Xia, J., Xue, Q. L., Harris, T. B., Purchase-Helzner, E., ... & Health ABC Study Group, F. (2013). Hearing loss and cognitive decline in older adults. *JAMA internal medicine*, *173*(4), 293-299.

Loizou, P. C. (1997, August). Signal processing for cochlear prosthesis: a tutorial review. In *Circuits and Systems*, *1997. Proceedings of the 40th Midwest Symposium on*, *2*, pp. 881-885). IEEE.

Lorenzi, C., Gilbert, G., Carn, H., Garnier, S., & Moore, B. C. (2006). Speech perception problems of the hearing impaired reflect inability to use temporal fine structure. *Proceedings of the National Academy of Sciences*, *103*(49), 18866-18869.

Lovelace, C. T., Stein, B. E., & Wallace, M. T. (2003). An irrelevant light enhances auditory detection in humans: a psychophysical analysis of multisensory integration in stimulus detection. *Cognitive brain research*, *17*(2), 447-453.

Lusk, L. G., & Mitchel, A. D. (2016). Differential gaze patterns on eyes and mouth during audiovisual speech segmentation. *Frontiers in psychology*, *7*, 52.

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*(4), 505-518.

MacDonald, K., Marchman, V., Fernald, A., & Frank, M. C. (2018). Adults and preschoolers seek visual information to support language comprehension in noisy environments. In *Oral presentation at the 40th Annual Meeting of the Cognitive Science Society*.

MacLeod, A., & Summerfield, Q. (1987). Quantifying the contribution of vision to speech perception in noise. *British journal of audiology*, *21*(2), 131-141.

Marin, S., Pouplier, M., & Harrington, J. (2010). Acoustic consequences of articulatory variability during productions of/t/and/k/and its implications for speech error research. *The Journal of the Acoustical Society of America*, *127*(1), 445-461.

Markham, D., & Hazan, V. (2002). The UCL Talker Database. *Speech, Hearing and Language: UCL Work in Progress*, *14*, 1–17.

Markham, D., & Hazan, V. (2004). The effect of talker-and listener-related factors on intelligibility for a real-word, open-set perception test. *Journal of Speech, Language, and Hearing Research*, *47*(4), 725-737.

Massaro, D. W. (1989). Testing between the TRACE model and the fuzzy logical model of speech perception. *Cognitive psychology*, *21*(3), 398-421.

Massaro, D. W., Cohen, M. M., & Smeele, P. M. (1996). Perception of asynchronous and conflicting visual and auditory speech. *The Journal of the Acoustical Society of America*, *100*(3), 1777-1786.

Mayer, J. L. (2017). The relationship between autistic traits and atypical sensory functioning in neurotypical and ASD adults: a spectrum approach. *Journal of autism and developmental disorders*, *47*(2), 316-327.

McCreery, R. W., Spratford, M., Kirby, B., & Brennan, M. (2017). Individual differences in language and working memory affect children's speech recognition in noise. *International journal of audiology*, *56*(5), 306-315.

McGrath, M., & Summerfield, Q. (1985). Intermodal timing relations and audio-visual speech recognition by normal-hearing adults. *The Journal of the Acoustical Society of America*, *77*(2), 678-685.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746.

Meredith, M. A., & Stein, B. E. (1986). Visual, auditory, and somatosensory convergence on cells in superior colliculus results in multisensory integration. *Journal of neurophysiology*, *56*(3), 640-662.

Middelweerd, M. J., & Plomp, R. (1987). The effect of speechreading on the speech-reception threshold of sentences in noise. *The Journal of the Acoustical Society of America*, *82*(6), 2145-2147.

Miles, J. H. (2011). Autism spectrum disorders—a genetics review. *Genetics in Medicine*, *13*(4), 278.

Miller, G. A. (1947). The masking of speech. *Psychological bulletin*, *44*(2), 105.

Miller, G. A., & Isard, S. (1963). Some perceptual consequences of linguistic rules. *Journal of Verbal Learning and Verbal Behavior, 2*(3), 217-228.

Millman, R. E., & Mattys, S. (2016). Auditory verbal working memory as a predictor of speech perception in modulated maskers in normal-hearing listeners. *Journal of speech, language, and hearing research*.

Mills, A. (1987). The development of phonology in the blind child, in Dodd, B. & Campbell, R. *Hearing by Eye: The Psychology of Lip-Reading* (pp.145–161). London: Lawrence Erlbaum Associates, Inc.

Moberly, A. C., Bates, C., Harris, M. S., & Pisoni, D. B. (2016). The Enigma of Poor Performance by Adults With Cochlear Implants. *Otology & neurotology: official publication of the American Otological Society, American Neurotology Society [and] European Academy of Otology and Neurotology*, *37*(10), 1522-1528.

Montgomery, A. A., & Jackson, P. L. (1983). Physical characteristics of the lips underlying vowel lipreading performance. *The Journal of the Acoustical Society of America*, *73*(6), 2134-2144.

Moon, I. J., & Hong, S. H. (2014). What is temporal fine structure and why is it important? *Korean journal of audiology*, *18*(1), 1.

Moore, B. C. (2008). The role of temporal fine structure processing in pitch perception, masking, and speech perception for normal-hearing and hearing-impaired people. *Journal of the Association for Research in Otolaryngology*, *9*(4), 399-406.

Moossavi, A., Etemadi, M., Javanbakht, M., Bakhshi, E., & Sharafi, M. A. (2016). Relationship between working memory capacity and speech perception in noise among children with cochlear implant. *Auditory and Vestibular Research*, *25*(4), 227-233.

Morrel-Samuels, P., & Krauss, R. M. (1992). Word familiarity predicts temporal asynchrony of hand gestures and speech. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(3), 615.

Munhall, K. G., Gribble, P., Sacco, L., & Ward, M. (1996). Temporal constraints on the McGurk effect. *Perception & psychophysics*, *58*(3), 351-362.

Nava, E., & Pavani, F. (2013). Changes in sensory dominance during childhood: Converging evidence from the Colavita effect and the sound-induced flash illusion. *Child Development*, *84*(2), 604-616.

Nelson, P. B., Jin, S. H., Carney, A. E., & Nelson, D. A. (2003). Understanding speech in modulated interference: Cochlear implant users and normal-hearing listeners. *The Journal of the Acoustical Society of America*, *113*(2), 961-968.

Nygaard, L. C., & Pisoni, D. B. (1998). Talker-specific learning in speech perception. *Perception & psychophysics*, *60*(3), 355-376.

O'Reilly, R.C. & Morlet, T. (September, 2016) Cochlear Implants. Retrieved from https://kidshealth.org/en/parents/cochlear.html

O'Connell, B. P., Hunter, J. B., & Wanna, G. B. (2016). The importance of electrode location in cochlear implantation. *Laryngoscope investigative otolaryngology*, *1*(6), 169-174.

Park, D. C., Lautenschlager, G., Hedden, T., Davidson, N. S., Smith, A. D., & Smith, P. K. (2002). Models of visuospatial and verbal memory across the adult life span. *Psychology and aging*, *17*(2), 299.

Peelle, J. E., & Sommers, M. S. (2015). Prediction and constraint in audiovisual speech perception. *Cortex*, *68*, 169-181.

Peters, R. W., Moore, B. C., & Baer, T. (1998). Speech reception thresholds in noise with and without spectral and temporal dips for hearing-impaired and normally hearing people. *The Journal of the Acoustical Society of America*, *103*(1), 577-587.

Picou, E. M., Ricketts, T. A., & Hornsby, B. W. (2011). Visual cues and listening effort: Individual variability. *Journal of Speech, Language, and Hearing Research*, *54*(5), 1416-1430.

Pimperton, H., Ralph-Lewis, A., & MacSweeney, M. (2017). Speechreading in deaf adults with cochlear implants: Evidence for perceptual compensation. *Frontiers in psychology*, *8*, 106.

Pisoni, D. B., Kronenberger, W. G., Chandramouli, S. H., & Conway, C. M. (2016). Learning and memory processes following cochlear implantation: the missing piece of the puzzle. *Frontiers in psychology*, *7*, 493.

Pollack, I. (1975). Auditory informational masking. *The Journal of the Acoustical Society of America*, *57*(S1), S5-S5.

Qin, M. K., & Oxenham, A. J. (2003). Effects of simulated cochlear-implant processing on speech reception in fluctuating maskers. *The Journal of the Acoustical Society of America*, *114*(1), 446-454.

Reisberg, D., Mclean, J., & Goldfield, A. (1987). Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli, in Dodd, B. & Campbell, R. *Hearing by Eye: The Psychology of Lip-Reading* (pp.97–114). London: Lawrence Erlbaum Associates, Inc.

Rennig, J., Wegner-Clemens, K., & Beauchamp, M. S. (2018). Face Viewing Behavior Predicts Multisensory Gain During Speech Perception. *bioRxiv*, 331306.

Robertson, I. H., Ward, T., Ridgeway, V., & Nimmo-Smith, I. (1994). *The test of everyday attention (TEA)*. Bury St Edmunds: Thames Valley Test Company.

Rönnberg, J., Lunner, T., Zekveld, A., Sörqvist, P., Danielsson, H., Lyxell, B., ... & Rudner, M. (2013). The Ease of Language Understanding (ELU) model: theoretical, empirical, and clinical advances. *Frontiers in systems neuroscience, 7*, 31.

Rosemann, S., Gießing, C., Özyurt, J., Carroll, R., Puschmann, S., & Thiel, C. M. (2017). The Contribution of Cognitive Factors to Individual Differences in Understanding Noise-Vocoded Speech in Young and Older Adults. *Frontiers in human neuroscience*, *11*, 294.

Rosen, S., Faulkner, A., & Wilkinson, L. (1999). Adaptation by normal listeners to upward spectral shifts of speech: Implications for cochlear implants. *The Journal of the Acoustical Society of America*, *106*(6), 3629-3636.

Rosen, S. M., Fourcin, A. J., & Moore, B. C. (1981). Voice pitch as an aid to lipreading. *Nature*, *291*(5811), 150.

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. *The Journal of the Acoustical Society of America, 133*(4), 2431-2443.)

Rosenblum, L. D. (2008). Speech perception as a multimodal phenomenon. *Current Directions in Psychological Science*, *17*(6), 405-409.

Rosenblum, L. D., Miller, R. M., & Sanchez, K. (2007). Lip-Read Me Now, Hear Me Better Later Cross-Modal Transfer of Talker-Familiarity Effects. *Psychological Science*, *18*(5), 392-396.

Ross, L. A., Saint-Amour, D., Leavitt, V. M., Javitt, D. C., & Foxe, J. J. (2006). Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cerebral Cortex*, *17*(5), 1147-1153.

Rothauser, E. H., Chapman, W. D., Guttman, N., Nordby, K. S., Silbiger, H. R., Urbanek, G. E., & Weinstock, M. (1969). IEEE recommended practice for speech quality measurements. IEEE Trans. *Audio Electroacoust, 17*(3), 225-246.

Rouger, J., Lagleyre, S., Démonet, J. F., Fraysse, B., Deguine, O., & Barone, P. (2012). Evolution of crossmodal reorganization of the voice area in cochlear-implanted deaf patients. *Human brain mapping*, *33*(8), 1929-1940.

Rouger, J., Lagleyre, S., Fraysse, B., Deneve, S., Deguine, O. & Barone, P. (2007). Evidence that cochlear-implanted deaf patients are better multisensory integrators. *Proceedings of the National Academy of Sciences of the United States of America, 104*, 7295–7300.

Rudner, M., Rönnberg, J., & Lunner, T. (2011). Working memory supports listening in noise for persons with hearing impairment. *Journal of the American Academy of Audiology*, *22*(3), 156-167.

Rumeau, C., Frère, J., Montaut-Verient, B., Lion, A., Gauchard, G., & Parietti-Winkler, C. (2015). Quality of life and audiologic performance through the ability to phone of cochlear implant users. *European Archives of Oto-Rhino-Laryngology*, *272*(12), 3685-3692.

Ruzich, E., Allison, C., Smith, P., Watson, P., Auyeung, B., Ring, H., & Baron-Cohen, S. (2015). Measuring autistic traits in the general population: a systematic review of the Autism-

Spectrum Quotient (AQ) in a nonclinical population sample of 6,900 typical adult males and females. *Molecular autism*, *6*(1), 2.

Sanchez, K., Dias, J. W., & Rosenblum, L. D. (2013). Experience with a talker can transfer across modalities to facilitate lipreading. *Attention, Perception, & Psychophysics*, *75*(7), 1359-1365.

Saunders, G. H., & Haggard, M. P. (1989). The clinical assessment of obscure auditory dysfunction--1. Auditory and psychological factors. *Ear and hearing*, *10*(3), 200-208.

Schoof, T., & Rosen, S. (2015). High sentence predictability increases the fluctuating masker benefit. *The Journal of the Acoustical Society of America*, *138*(3), EL181-EL186.

Scott, S. K., Blank, C. C., Rosen, S., & Wise, R. J. (2000). Identification of a pathway for intelligible speech in the left temporal lobe. *Brain*, *123*(12), 2400-2406.

Sekiyama, K., Kanno, I., Miura, S., & Sugita, Y. (2003). Auditory-visual speech perception examined by fMRI and PET. *Neuroscience research*, *47*(3), 277-287.

Shams, L., Kamitani, Y., & Shimojo, S. (2002). Visual illusion induced by sound. *Cognitive Brain Research*, *14*(1), 147-152.

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, *270*(5234), 303-304.

Sheffert, S. M., Lachs, L., & Hernandez, L. R. (1996). The Hoosier audiovisual multi-talker database. *Research on spoken language processing progress report*, *21*, 578-583.

Shepherd, D. C., DeLavergne, R. W., Frueh, F. X., & Clobridge, C. (1977). Visual-neural correlate of speechreading ability in normal-hearing adults. *Journal of Speech, Language, and Hearing Research*, *20*(4), 752-765.

Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends in cognitive sciences*, *12*(5), 182-186.

Simpson, S. A., & Cooke, M. (2005). Consonant identification in N-talker babble is a nonmonotonic function of N. *The Journal of the Acoustical Society of America*, *118*(5), 2775-2778.

Skuk, V. G., Palermo, R., Broemer, L., & Schweinberger, S. R. (2017). Autistic Traits are Linked to Individual Differences in Familiar Voice Identification. *Journal of autism and developmental disorders*, 1-21.

Smith, E. G., & Bennetto, L. (2007). Audiovisual speech integration and lipreading in autism. *Journal of Child Psychology and Psychiatry*, *48*(8), 813-821.

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.

Souza, P., & Rosen, S. (2009). Effects of envelope bandwidth on the intelligibility of sine- and noise-vocoded speech. *The Journal of the Acoustical Society of America*, *126*(2), 792-805.

Souza, P., Gehani, N., Wright, R., & McCloy, D. (2013). The advantage of knowing the talker. *Journal of the American Academy of Audiology*, *24*(8), 689-700.

Spivak, L. G., & Waltzman, S. B. (1990). Performance of cochlear implant patients as a function of time. *Journal of Speech, Language, and Hearing Research*, *33*(3), 511-519.

Stacey, P. C., & Summerfield, Q. A. (2007). Effectiveness of computer-based auditory training in improving the perception of noise-vocoded speech. *The Journal of the Acoustical Society of America*, *121*(5), 2923-2935.

Stacey, P. C., Kitterick, P. T., Morris, S. D., & Sumner, C. J. (2016). The contribution of visual information to the perception of speech in noise with and without informative temporal fine structure. *Hearing research*, *336*, 17-28.

Stein, B. E., Stanford, T. R., & Rowland, B. A. (2014). Development of multisensory integration from the perspective of the individual neuron. *Nature Reviews Neuroscience*, *15*(8), 520.

Stevenson, R. A., Segers, M., Ferber, S., Barense, M. D., Camarata, S., & Wallace, M. T. (2016). Keeping time in the brain: Autism spectrum disorder and audiovisual temporal processing. *Autism Research*, *9*(7), 720-738.

Stevenson, R. A., Segers, M., Ncube, B. L., Black, K. R., Bebko, J. M., Ferber, S., & Barense, M. D. (2018). The cascading influence of multisensory processing on speech perception in autism. *Autism*, *22*(5), 609-624.

Stevenson, R. A., Sheffield, S. W., Butera, I. M., Gifford, R. H., & Wallace, M. T. (2017). Multisensory integration in cochlear implant recipients. *Ear and hearing*, 38(5), 521-538.

Stevenson, R. A., Zemtsov, R. K., & Wallace, M. T. (2012). Individual differences in the multisensory temporal binding window predict susceptibility to audiovisual illusions. *Journal of Experimental Psychology: Human Perception and Performance*, *38*(6), 1517.

Stickney, G. S., Zeng, F. G., Litovsky, R., & Assmann, P. (2004). Cochlear implant speech recognition with speech maskers. *The Journal of the Acoustical Society of America*, *116*(2), 1081-1091.

Stone, J. V., Hunkin, N. M., Porrill, J., Wood, R., Keeler, V., Beanland, M., ... & Porter, N. R. (2001). When is now? Perception of simultaneity. *Proceedings of the Royal Society of London B: Biological Sciences*, *268*(1462), 31-38.

Strelcyk, O., & Dau, T. (2009). Relations between frequency selectivity, temporal fine-structure processing, and speech reception in impaired hearing. *The Journal of the Acoustical Society of America*, *125*(5), 3328-3345.

Stropahl, M., & Debener, S. (2017). Auditory cross-modal reorganization in cochlear implant users indicates audio-visual integration. *NeuroImage: Clinical*, *16*, 514-523.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The 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 Dodd, B. & Campbell, R. *Hearing by Eye: The Psychology of Lip-Reading* (pp.3–52). London: Lawrence Erlbaum Associates, Inc.

Summerfield, Q. (1992). Lipreading and audio-visual speech perception. *Phil. Trans. R. Soc. Lond. B, 335*(1273), 71-78.

Talsma, D., Senkowski, D., Soto-Faraco, S., & Woldorff, M. G. (2010). The multifaceted interplay between attention and multisensory integration. *Trends in cognitive sciences, 14*(9), 400-410.

Tanaka, J. W., & Sung, A. (2016). The “eye avoidance” hypothesis of autism face processing. *Journal of autism and developmental disorders*, *46*(5), 1538-1552.

Theeuwes, J. (1991). Exogenous and endogenous control of attention: The effect of visual onsets and offsets. *Perception & psychophysics*, *49*(1), 83-90.

Turner, C. W., Gantz, B. J., Vidal, C., Behrens, A., & Henry, B. A. (2004). Speech recognition in noise for cochlear implant listeners: benefits of residual acoustic hearing. *The Journal of the Acoustical Society of America*, *115*(4), 1729-1735.

Tye-Murray, N., Hale, S., Spehar, B., Myerson, J., & Sommers, M. S. (2014). Lipreading in school-age children: the roles of age, hearing status, and cognitive ability. *Journal of Speech, Language, and Hearing Research*, *57*(2), 556-565.

Tye-Murray, N., Sommers, M., Spehar, B., Myerson, J., & Hale, S. (2010). Aging, audio-visual integration, and the principle of inverse effectiveness. *Ear and hearing*, *31*(5), 636.

Tye-Murray, N., Spehar, B., Myerson, J., Hale, S., & Sommers, M. (2016). Lipreading and audiovisual speech recognition across the adult lifespan: Implications for audiovisual integration. *Psychology and aging*, *31*(4), 380.

Tyler, R. S., Gantz, B. J., Woodworth, G. G., Fryauf-Bertschy, H., & Kelsay, D. M. (1997). Performance of 2-and 3-year-old children and prediction of 4-year from 1-year performance. *The American journal of otology*, *18*(6 Suppl), S157-9.

Unsworth, N., Heitz, R. P., Schrock, J. C., & Engle, R. W. (2005). An automated version of the operation span task. *Behavior research methods*, *37*(3), 498-505.

Vabalas, A., & Freeth, M. (2016). Brief report: Patterns of eye movements in face to face conversation are associated with autistic traits: Evidence from a student sample. *Journal of autism and developmental disorders*, *46*(1), 305-314.

Van Engen, K. J., Xie, Z., & Chandrasekaran, B. (2017). Audiovisual sentence recognition not predicted by susceptibility to the McGurk effect. *Attention, Perception, & Psychophysics*, *79*(2), 396-403.

Van Wassenhove, V., Grant, K. W., & Poeppel, D. (2007). Temporal window of integration in auditory-visual speech perception. *Neuropsychologia*, *45*(3), 598-607.

Vatakis, A., & Spence, C. (2006). Audiovisual synchrony perception for speech and music assessed using a temporal order judgment task. *Neuroscience letters*, *393*(1), 40-44.

Vatikiotis-Bateson, E., Eigsti, I. M., Yano, S., & Munhall, K. G. (1998). Eye movement of perceivers during audiovisualspeech perception. *Perception & psychophysics*, *60*(6), 926-940.

Vermeire, K., Brokx, J. P., Wuyts, F. L., Cochet, E., Hofkens, A., & Van de Heyning, P. H. (2005). Quality-of-life benefit from cochlear implantation in the elderly. *Otology & Neurotology*, *26*(2), 188-195.

von dem Hagen, E. A., & Bright, N. (2017). High autistic trait individuals do not modulate gaze behaviour in response to social presence but look away more when actively engaged in an interaction. *Autism Research*, *10*(2), 359-368.

Vroomen, J., & Keetels, M. (2010). Perception of intersensory synchrony: a tutorial review. *Attention, Perception, & Psychophysics*, *72*(4), 871-884.

Wagner, A. E., Toffanin, P., & Başkent, D. (2016). The timing and effort of lexical access in natural and degraded speech. *Frontiers in psychology*, *7*, 398.

Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science*, *167*(3917), 392-393.

Weatherholtz, K., & Jaeger, T. F. (2016). Speech perception and generalization across talkers and accents. In *Oxford Research Encyclopedia of Linguistics*.

Wechsler, D. (1991). *Wechsler intelligence scale for children*. The Psychological Corporation. Texas: Pearson.

Wechsler, D. (2008). *Wechsler Intelligence Scale–Fourth Edition (WAIS-IV): Technical and interpretive manual*. Texas: Pearson

Wertheimer, M. (1961). Psychomotor coordination of auditory and visual space at birth. *Science*, *134*(3491), 1692-1692.

Whitmal III, N. A., Poissant, S. F., Freyman, R. L., & Helfer, K. S. (2007). Speech intelligibility in cochlear implant simulations: Effects of carrier type, interfering noise, and subject experience. *The Journal of the Acoustical Society of America*, *122*(4), 2376-2388.

Winneke, A. H., & Phillips, N. A. (2011). Does audiovisual speech offer a fountain of youth for old ears? An event-related brain potential study of age differences in audiovisual speech perception. *Psychology and aging*, *26*(2), 427.

Yi, A., Wong, W., & Eizenman, M. (2013). Gaze patterns and audiovisual speech enhancement. *Journal of Speech, Language, and Hearing Research*, *56*(2), 471-480.

Zeng, F. G., Rebscher, S. J., Fu, Q. J., Chen, H., Sun, X., Yin, L., ... & Yang, B. (2015). Development and evaluation of the Nurotron 26-electrode cochlear implant system. *Hearing research*, *322*, 188-199.