THE INFLUENCE OF VISUAL INFORMATION ON THE PERCEPTION OF AUDITORY SPEECH IN QUIET AND NOISE

JEMAINE ELEANOR STACEY

A dissertation submitted to the faculty of

Nottingham Trent University

In partial fulfilment for the degree of

Doctor of Philosophy

September 2019

Copyright Statement

Abstract

Audio-visual (AV) integration involves the combining of auditory and visual information which is often required for everyday face to face communication. Speech perception becomes difficult in situations when it is harder to hear the voice of the speaker. When the ability to identify speech in noise is reduced, people with normal hearing improve with the addition of visual information; when they can see the talker's face (Sumby & Pollack, 1954). Exactly how visual information is used in background noise is not well understood. The goal of the thesis was to understand the influence of visual information on auditory speech perception using a famous measure of AV integration (The McGurk effect). Four experiments are reported which aimed to a) explore the use of the McGurk effect as a measure of AV integration, b) understand the influence of visual information in quiet and noise, and how auditory and visual information interact when one or both of the modalities is degraded, and c) provide insight into theories of AV integration through using behavioural measures. The main findings were that 1) instances of the McGurk effect are influenced by the type of task used, and vary according to different stimuli and participants, 2) The McGurk effect can still be perceived even when the visual stimulus is highly degraded although the illusion decreases as visual blur increases, 3) fixating the mouth is not necessary for perceiving the McGurk effect, 4) Visual benefit increases as the clarity of the visual stimulus increases. Overall, the findings suggest that visual information is of most benefit when it is clear, looking at the mouth is not necessary for AV integration in quiet but increases the likelihood of successful integration when speech is presented in auditory noise.

Financial support

The PhD was funded by the Vice Chancellor Scholarship at Nottingham Trent University.

Declaration

This thesis comprises the candidate’s own work and has not been submitted to this or any University for a degree. All aspects of the thesis were completed by the candidate.

Publications/conference proceedings

Experiment 1

Stacey, J. E., Howard, C., Mitra, S. & Stacey, P. (2016). Manipulating the McGurk effect. Oral presentation: *Psychology Postgraduate affairs Group* (PsyPAG); York, UK.

Experiments 2 & 3

Stacey, J. E., Howard, C., Mitra, S. & Stacey, P. (In press). Audio-visual integration in noise: Influence of auditory and visual stimulus degradation on eye movements and perception of the McGurk effect. Attention, Perception and Psychophysics.

Stacey, J. E., Howard, C., Mitra, S. & Stacey, P. (2017). The relationship between eye movements and the McGurk effect when stimuli are presented in noise. Poster presented at: *International Multisensory Research Forum*; May 19-22; Nashville, TN, USA.

Stacey, J. E. (2017). Auditory-visual integration in noise: A brief review. *Cognitive Psychology Bulletin*, 2(1), 30-32.

Contents

Figures

Tables

Acknowledgements

I would like to thank my supervisors, thank you Paula Stacey for keeping me going throughout my PhD with enthusiasm and positivity, to Christina Howard for emotional support, and to Suvo Mitra for imparting all your knowledge and wisdom.

I'm grateful to my partner David for putting up with me, making me dinner when I was exhausted and for always believing in me.

Thank you to all my friends who have been my cheerleaders throughout this whole process including; Sarah Gardner, Andrew MacKenzie, Chloe Wider, Harriet Smith, Lydia Harkin, Lucy Justice & Suzanne Neumann. My PsyPAG family, including Catherine Talbot, Scott Orr, Liam Knox, Ryc Aquino, Gareth Richards & Tommy Van Steen.

A huge thank you to the technicians, Joshua Baker , Roy Barson, Steve Scott and Ben Thompson for technical support (and calming me down when everything went wrong).

Thank you to the NTU psychology department who have been friendly and supportive throughout this process.

Chapter 1: Audio-visual integration

1.1 Non-speech audio-visual illusions

In everyday life multisensory information from our environment helps us form a coherent percept of the world. In particular, visual and auditory information often convey consistent information, an example is seeing someone walking and simultaneously hearing their footsteps. However, perceptual illusions can occur when incongruent information from different modalities is presented simultaneously. This is demonstrated in the sound-induced flash illusion in which viewers perceive a single flash of light as a double flash if it coincides with two auditory beeps (Shams, Kamitani & Shimojo, 2000). Two flashes can also be perceived as a single flash if a single beep is presented, this is termed the fusion effect as it appears as though the two flashes have 'fused' into one flash (Anderson, Tiippana & Sams, 2004). Similarly, in the cross-bounce illusion when two circles cross whilst a beep is simultaneously presented this makes the circles appear as though they have bounced (Sekular, Sekular & Lau, 1997). These illusions demonstrate how auditory stimuli can influence visual perception. Visual information can also constrain auditory perception, for example, when watching a film, the sound originates from the cinema/television loudspeakers which are not at the same precise location in space as the picture of the mouth on the screen, yet the sound appears to originate from the mouth of the talker. When visual cues determine the perceived location of an auditory stimulus this is termed the ventriloquism effect (Howard & Templeton, 1966). Collectively, these illusions show that both vision and audition can influence the other under different circumstances.

These observations have led to discussions about the nature of the sensory pathways in the brain and how multisensory information combines to produce unitary percepts (theories will be covered in Section 1.5). Evidence from neuropsychological research shows that visual brain regions can respond to auditory stimuli and vice versa (Shams, Kamitani & Shimojo, 2001). Shams et al. (2001) found that the sound-induced flash illusion resulted in activity in visual parts of the brain that would be activated if a real flash were perceived suggesting that vision is influenced by auditory information in brain regions responsible for visual processing. This suggests that sensory pathways can be cross-modal as they can be influenced by other modalities (Shimojo & Shams, 2001; Shams & Kim, 2010).

1.2 The McGurk effect

The McGurk effect (McGurk & MacDonald, 1976) is a famous phenomenon that has been used in over 40 years of research (Rosenblum, 2019) to investigate AV integration in speech perception. This phenomenon demonstrates how information from the auditory and visual modalities is combined to produce a unitary percept (Tiippana, 2014). This illusion occurs when incongruent auditory and visual syllables are presented simultaneously resulting in an illusory percept. For example, hearing a voice utter the syllable /ba:/ (auditory /ba:/ = A_{BA}) whilst viewing lip movements uttering /ga:/ (visual /ga:/ = V_{GA}) has the effect that listeners perceive a different syllable to that of the auditory or visual syllable e.g. /da:/ or /θa:/. As this results in the perception of a third or different syllable, this is termed a fusion response. This occurs because the visual information influences the auditory information causing the listener to perceive something other than what was said. This syllable combination has been reported to produce the illusion the most consistently compared to other syllable combinations (McGurk & MacDonald, 1976). Different syllables also produce the illusion for example A_{BA} and V_{VA} results in /va:/ which suggests visual dominance. Furthermore, the combination of A_{GA} and V_{BA} produces a blend of two syllables for example /bga:/ (McGurk & MacDonald, 1976). Massaro and Cohen (1993) manipulated the degree of synchrony with which auditory and visual information were presented together. They presented vowels and consonant-vowel syllables (e.g. ba), it was found that for consonants AV integration occurred regardless of the asynchrony. Presenting an A_{DA} with V_{BA} resulted in a ‘b-da’ whereas presenting an auditory vowel with a visual vowel rarely resulted in integration suggesting that incongruent consonants are needed to produce increased instances of the illusion.

In a review, Tiippana (2014) highlights the issue of defining the McGurk effect and states that it may be difficult to gauge the prevalence of the McGurk effect due to the different definitions utilised by different studies. Some consider the fusion response to be the true McGurk effect because it produces a third or different syllable to that of the auditory or visual syllables. However, this definition does not include other syllable combinations which produce blends of both the auditory and visual information e.g. /bga:/. Tiippana (2014) advocates the classical definition in which an individual reports anything other than the auditory percept, as this definition accounts for all McGurk syllable combinations (McGurk & MacDonald, 1976), this means that all incorrect responses to the auditory syllable are counted as an illusory effect.

The McGurk effect is generally considered a robust illusion and has been replicated across different languages although different languages produce the effect to different extents (Massaro, Cohen, Gesi, Heredia, & Tsuzaki, 1993; Sekiyama. & Tokhura, 1991). It is stable across time, as studies show that individuals who were tested initially and then again two months later (Strand, Cooperman, Rowe and Simenstad, 2014) and one year later (Basu-Mallick, Magnotti & Beauchamp, 2015) were found to perceive the McGurk effect to the same extent at both time points.

1.2.1 Individual differences in the McGurk effect

Although considered robust due to the numerous replications of the illusion, the McGurk effect has been found to vary substantially across individuals, with some individuals never perceiving the McGurk effect and some people perceiving it on every trial. Studies have reported different estimates of McGurk susceptibility ranging from 0-100% (Basu-Mallick et al., 2015; Nath & Beauchamp, 2012) and 1-91% (Benoit, Raij, Lin, Jääskeläinen, & Stufflebeam, 2010) of trials. It is not well established why individuals vary in their ability to perceive the McGurk effect.

Strand et al. (2014) wanted to explore factors which could account for individual differences in the McGurk effect. The participants completed a lip-reading task which required them to identify silent videos of consonants and words. There were two separate tasks involving incongruent stimuli, one in which the participants reported what they heard and another where they reported whether the auditory and visual information were congruent or not. It was found that lip-reading was positively related to McGurk perception. However, the findings are contradictory as they also suggest that good lip-readers were also better at detecting incongruent AV information which should result in fewer McGurk responses. This suggests that more proficient lip-readers were able to utilise visual information in different ways depending on the task. Brown et al. (2018) found that individuals who are better at lip-reading also perceive the McGurk effect more often. This finding has also been observed for people with hearing impairments (Grant & Seitz, 1998). Brown et al. (2018) also reported that other cognitive abilities including processing speed, attentional control and working memory were not related to McGurk effect perception. Overall, these results suggest that good lip-readers are better at extracting visual information from the stimulus and are therefore more susceptible to the influence of visual information which results in perceiving the McGurk effect more often.

Some of the variability in the McGurk effect can also be accounted for by methodological factors. It is difficult to determine the true prevalence of the McGurk effect as studies may only report data from people who perceive the McGurk effect meaning it is not known how many participants in the sample did not perceive the McGurk effect at all. Therefore, results may not be representative of the population. Furthermore, different studies also use different stimuli which can influence the amount the McGurk effect is perceived as stimulus properties vary including; the talker used in the video, size of the talker’s face, video quality and synchronisation of the syllables (Basu-Mallick et al., 2015). Methodological factors that influence the McGurk effect will be explored further in Experiment 1 (Chapter 3).

1.3 Weightings of visual and auditory information

To further understanding of how our senses interact it is useful to determine what influences the perception of AV illusions. In a review, Shams and Kim (2010) point out that vision is often viewed as the dominant sense, this is evidenced in multiple papers in which humans are referred to as ‘visual animals’ (e.g. Shimojo & Shams, 2001). However, the factors which determine which sense dominates are not fully understood. Whether audition or vision dominates is context dependent (Walker & Scott, 1981) and can depend on the demands of the task (Robinson, Chandra & Sinnett, 2016). Sound has temporal properties which means that audition dominates in temporal tasks, Walker and Scott (1981) found that when an auditory stimulus (tone) was presented separately from a visual stimulus (light) of the same duration the tone was perceived as longer than the light, and when the tone and light were presented simultaneously the light was perceived as the same duration as the tone when it was presented alone. This suggests that temporal judgements of visual stimuli were being influenced by the duration of auditory stimuli. In contrast, vision contains spatial information and therefore dominates in spatial tasks, this is evident in the ventriloquist effect (Howard & Templeton, 1966) as vision alters the perceived auditory location.

Other factors also affect whether visual or auditory information dominates. Robinson et al. (2016) presented the participants with an auditory (tone), visual (shape) and bimodal oddballs (tone + shape), they found that when only one response key was required for all three stimulus types audition was dominant, however increasing the response options to three resulted in a switch from auditory to visual dominance. The finding that additional demands on attention resulted in visual dominance suggests that the participants may have a bias for visual information over auditory or bimodal information. Visual dominance has also been found to increase with age (Hirst, Stacey, Cragg, Stacey & Allen, 2018; Sekiyama, Soshi & Sakamoto, 2014). Moreover, the modality appropriate hypothesis suggests that the most reliable modality is the one that dominates (Welch & Warren, 1980). Auditory or visual dominance can depend on the weighted reliability of information from each sense (Ernst & Bülthoff, 2004; Witten & Knudsen, 2005). When faced with the task of understanding speech in quiet listening conditions, audition is the dominant sense as speech can be easily identified from auditory information alone (Gatehouse & Gordon, 1990; Shannon, Zeng, Kamath, Wygonski & Ekelid, 1995). In contrast, it is very difficult to understand speech from visual information only (Bernstein & Liebenthal, 2014). However, during AV speech perception, if information in one modality is degraded this can shift sensory dominance to the more reliable sense and in turn influence AV integration. For example, trying to understand someone speaking in a noisy room may result in more reliance on the visual information (Sumby & Pollack, 1954).

1.4 Audiovisual integration of speech stimuli vs non-speech stimuli

It should be noted that the integration of auditory and visual information of non-speech stimuli may be different from the integration of auditory and visual information for speech perception for several reasons. Firstly, naturalistic speech which occurs in everyday conversation includes a social aspect of conversing with another talker whereas this is absent in non-speech stimuli. Second, individuals perform better on tasks if they believe they are being presented with AV speech stimuli compared to non-speech stimuli (Remez, Rubin, Pisoni & Carrell, 1981; Tuomainen, Andersen, Tiippana & Sams, 2005). Remez et al. (1981) presented one group of participants with sine wave speech and asked them to describe the stimuli, the participants reported that the stimuli were computer generated sounds (beeps & whistles). Another group were told that the stimuli were speech and were subsequently able to correctly identify more stimuli than the naïve group. Third, individuals are better at integrating AV speech information compared to non-speech information. This is evidenced by pluck-bow stimuli which were designed to be the non-speech equivalent of the McGurk effect (Saldaña & Rosenblum, 1993) Pluck-bow stimuli are comprised of plucks and bows from a cello, these stimuli elicited an illusion whereby the participants misjudged whether the auditory stimulus was a pluck or a bow if it was presented with incongruent visual information. Saldaña and Rosenblum (1993) compared McGurk syllables with AV videos of pluck-bow stimuli and found that McGurk syllables produced a stronger illusory effect than pluck-bow stimuli. Taken together, these results may suggest that AV speech is processed differently to AV non-speech information. Tuomainen et al. (2005) argue that the auditory and visual components of AV non-speech are processed separately whereas speech is combined and forms a unitary percept, resulting in more accurate identification of speech stimuli. However, exactly how AV speech is processed and the timing of AV integration has been debated.

1.5. Theoretical explanations of the McGurk effect

This section will now outline key theories of speech perception which relate to the McGurk effect. Traditional auditory theories such as that of Diehl and colleagues (e.g. Diehl, 1987) place an emphasis on the auditory signal alone and potentially underestimate the role of visual information in speech perception. Several theories are outlined which aid in understanding the influence of visual information, each theory also has a hypothesis in relation to the McGurk effect, and proposes when in time AV integration occurs.

1.5.1 The fuzzy logic model of perception (late integration)

The fuzzy logic model of perception (FLMP; Massaro & Oden, 1980) is classed as an auditory model of speech perception as an emphasis is placed on deciphering the acoustic signal. The FLMP proposes that speech perception is not specific to humans and there is no distinction between the processing of speech and other sounds. The theory outlines three key steps involved in speech perception, the first is feature evaluation which involves analysis of the properties of the auditory signal. These features are then compared with prototypes held in memory (prototype matching). Finally, the information about the features is matched with the most relevant prototype and the stimulus is identified (pattern classification). The McGurk effect can be explained by the FLMP in terms of this decision making process. Massaro and Cohen (1983) conducted several experiments using AV stimuli which were on a continuum from ba to da, by using this method they were able to manipulate the ambiguity of the syllable. RT was used to measure the speed of decision making and it was found that RT was slower when the auditory and visual information was incongruent compared to congruent. This reflects the longer processing time involved with resolving the ambiguity. These findings support the FLMP and suggest that each modality is processed separately but in parallel and that integration of the auditory and visual modalities occurs later in time (Massaro, 1987; Massaro & Cohen, 1983). Further research supporting this theory is outlined in Chapter 4.

1.5.2 Revised Motor Theory (early integration)

The motor theory of speech perception (Liberman, 1957), later updated to the Revised Motor Theory (RMT; Liberman & Mattingly, 1985) suggests that the motor commands (intended gestures) necessary for producing speech (e.g. lip movements) are also utilised for perceiving speech. This account places an emphasis on identifying the intended gestures of the talker rather than the auditory signal. An example is highlighted in the way different phonemes map onto visemes (outlined in Chapter 1 Section 1.8.1). As visemes enable the identification of speech, the intended gesture of the talker is identified followed by the phoneme. Direct realist theory (DRT; Fowler, 1981) is also consistent with the view that speech perception is achieved through identifying gestures however, in contrast to RMT, DRT purports that the literal articulations of the talker are sufficient for speech perception rather than related motor commands (intended gestures).

A related concept to RMT is analysis by synthesis (AbyS; Halle & Stevens, 1962) which describes an internal model of speech perception which involves pattern matching between a specific set of rules for producing speech and incoming speech. This synthesizer is considered innate, Liberman and Mattingly (1985) cite the evolution of the vocal tract in humans as evidence for this, this also means that speech perception is specific to humans.

RMT purports that speech perception cannot be explained by theoretical accounts of sound perception in general as the relationship between the auditory signal and related gestures is unique to speech, therefore, speech is considered 'special' although this idea is controversial and has been widely debated. Galantucci, Fowler and Turvey, (2006) suggest that this idea is open to interpretation and therefore difficult to test. DRT (Fowler, 1981) also rejects the idea that speech specific mechanisms are required for speech perception and suggests that there are universal mechanisms for perception utilised in other domains for example; visual object perception (Fowler, 1996).

The RMT explains the illusion arising from McGurk syllables as evidence that speech perception involves deciphering gestures. In their example, A_{BA} with V_{VA} results in the percept /va:/ because the emphasis is placed on the intended gesture of the talker (Liberman & Mattingly, 1985) which is more easily extracted from the visual information. Research which suggests that people benefit from being able to see a talker's face when listening in auditory noise (e.g. Sumby & Pollack, 1954) also supports this claim (a review of this research is provided in Section 1.9). Overall, this theory suggests that speech is not considered bimodal and that speech perception relies on the convergence of auditory and visual information early on to produce the intended gesture.

1.5.3 Models of AV integration

At what point in time auditory and visual information are integrated is unclear, models of AV integration operationalise the aforemetioned theories and describe how auditory and visual information combine. The models can be summarized into three classes outlined in Figure 1.1 (Peelle & Sommers, 2015). Late integration models propose that auditory and visual cues are processed separately before integration. Early integration models suggest that auditory and visual cues are integrated during perception and are at no point represented separately. A limitation of early and late models is that they describe AV integration as unidirectional and linear (van Wassenhove, 2013). Peelle and Sommers (2015) advocate the idea of a multistage model which suggests that auditory and visual cues are processed at both an earlier and later stage (Peelle & Sommers, 2015).

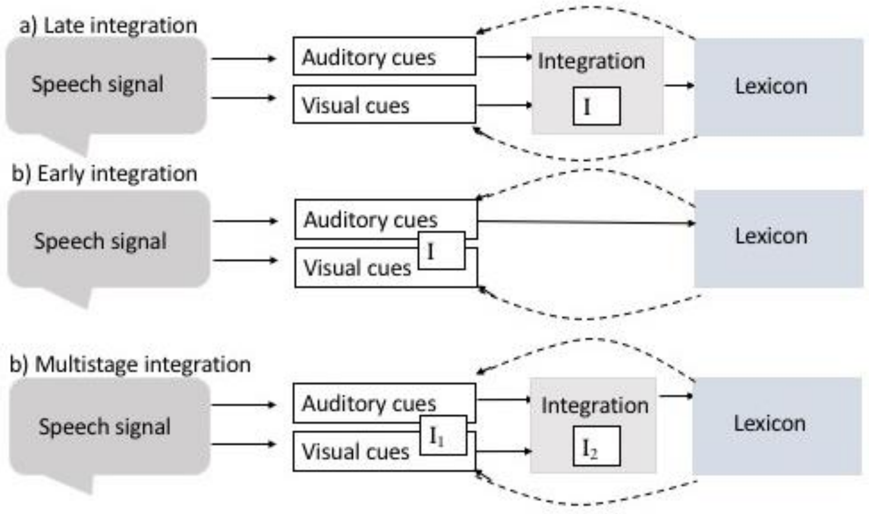


Figure 1.1 Late, early and multistage models of integration. a) the late integration model describes AV integration as occurring after separate auditory and visual information is processed, b) the early integration model suggests that integration occurs immediately during perception, and c) the multistage model suggests integration can occur both earlier and later (Peelle & Sommers, 2015).

At present, there is no agreement on which model best describes AV integration; these models can be assessed using incongruent stimuli and reaction time (See Chapter 3).

1.6 What can the McGurk effect tell us about AV integration in noise?

Perhaps the most useful application of the McGurk effect is to try and understand AV integration in noise. In quiet, AV integration is not always needed for successful speech perception as either modality can potentially be used whereas in noise, AV integration may provide an advantage for speech perception (Marques, Lapenta, Costa & Boggio, 2016). Sekiyama and Tokhura (1991) presented Japanese participants with AV McGurk syllables in quiet and in auditory noise. In quiet listening conditions, some participants did not perceive the Japanese McGurk effect at all whereas for others it was very minimal. An important finding was that when auditory noise was added, the proportion of illusory syllables evoked by incongruent stimuli increased. This indicates that individuals focused more on the visual information as this was more reliable which in turn resulted in increased instances of the McGurk effect.

The McGurk effect in noise has also been used to compare AV integration across different ages. In order to measure how well people can understand speech in background noise, the Signal-to-Noise ratio (SNR) at which people can perform is often measured. Positive SNRs indicate that the target speech is louder than the background noise, whereas negative SNRs indicate that the target speech is quieter than the background noise. Sekiyama, Soshi and Sakamoto (2014) presented congruent AV syllables and incongruent syllables in auditory noise at four SNRs from 0 dB to 18 dB increasing in steps of 6 dB. The trials in which the participants perceived the McGurk effect were subtracted from congruent trials to give a measure of visual benefit. They found that older adults received more visual benefit than younger adults across all noise conditions and that this increased as auditory noise increased. Hirst, Stacey, Cragg, Stacey & Allen (2018) compared children and adults using incongruent stimuli presented in quiet and four SNRs (-2 dB, -8 dB, -14 dB & -20 dB). They found that young children needed more auditory noise to increase McGurk responses compared to adults and older children. This suggests that the amount of visual influence received from viewing a talker's face increases with age. Studies like this are useful as they demonstrate how visual information can impact speech perception in difficult listening situations, and across the life span.

1.7 Validity of the McGurk effect as a measure of AV integration

It is important to note that the validity of the McGurk effect as a measure of AV integration has been questioned in recent years (Alsius, Paré & Munhall, 2017; Van Engen, Xie & Chandrasekaran, 2017). There is an underlying assumption in the literature that individuals who perceive the McGurk effect more often (strong perceivers) would also be more accurate at identifying congruent speech in noise compared to those who perceive the McGurk effect less often, because strong perceivers would be better at integrating information. However, recent research (Van Engen, Xie & Chandrasekaran, 2017) found that when sentences and incongruent stimuli were presented in noise (multi-talker babble) visual benefit for sentences was not predicted by the McGurk effect. However, the sample size used in this study was small and therefore may be underpowered. In a review, Alsius, Paré and Munhall (2017) suggest that caution should be taken when claiming that the McGurk effect is comparable to AV integration during everyday conversation. The authors (Alsius et al., 2017) highlight several key differences between congruent speech stimuli and incongruent stimuli in that the subjective experience of the McGurk effect may be different to that of everyday speech. Brancazio (2004) comapred congruent syllables with incongruent syllables and found that incongruent stimuli were rated as inferior examples of syllable categories. In addition, when the auditory and visual components of congruent speech stimuli and incongruent stimuli were temporally offset, incongruent stimuli were judged as more asynchronous (van Wassenhove et al., 2007). These findings are unsurprising, as the auditory and visual information are incongruent and provide conflicting information therefore subjective judgements about incongruent stimuli are expected to differ compared to congruent speech stimuli. Moreover, both AV congruent speech stimuli and AV incongruent stimuli provide phonetic context, whereas congruent speech stimuli also provides information about phonological and lexical constraints, semantic context and syntax (Van Engen, Dey, Sommers & Peelle, in press).

Despite these differences, there is evidence that McGurk effect perception and the identification of congruent speech stimuli share the same mechanism for AV integration. Grant and Seitz (1998) found that the frequency of the McGurk effect increased as visual benefit increased, for both consonants and sentences. This suggests that there is a relationship between integrating auditory and visual information in congruent AV speech, and the integration involved in perceiving the McGurk effect. Therefore, it would appear that the McGurk effect may provide a valuable measure for testing AV integration. Further evidence is provided by the finding that the higher incidence of McGurk perception when individuals are better at lip-reading (Strand et al., 2014). Cochlear-implant users experience an enhanced McGurk effect compared to NH listeners suggesting that they are more influenced by visual information (Rouger, Fraysse, Deguine & Barone, 2008). These results demonstrate how an individual's use of visual information can influence AV integration as measured by the McGurk effect. Furthermore, the McGurk effect provides a useful paradigm for understanding the influence of visual information when speech is presented in noise (outlined in Section 1.6).

The McGurk effect also has an advantage over speech in noise tasks which use sentences as the short nature of the stimuli means more trials can be used (more power), participants may experience less fatigue as a result, and reaction time can be easily measured to assess the timing of AV integration.

Due to the popularity of the McGurk effect, evidenced by numerous citations, it is often cited as a useful measure of AV integration (e.g. Tiippanna, 2014), therefore further work is needed to establish the relationship between the McGurk effect and everyday speech. If a relationship cannot be established between McGurk effect perception and everyday speech, then the McGurk effect will need to be reconceptualised in order to move forward.

Overall, the findings from experiments using the McGurk effect are informative for ascertaining how individuals use visual information and integrate incongruent AV speech. Whether or not these findings can be generalised to AV congruent speech is still an open question and further research is required to examine the McGurk effect in relation to other measures of AV integration. This research would have important implications, as the findings would help to determine the validity of the McGurk effect as a measure of AV integration.

1.8 Visual Speech perception: what can be understood from visual information alone?

1.8.1 Silent lip-reading

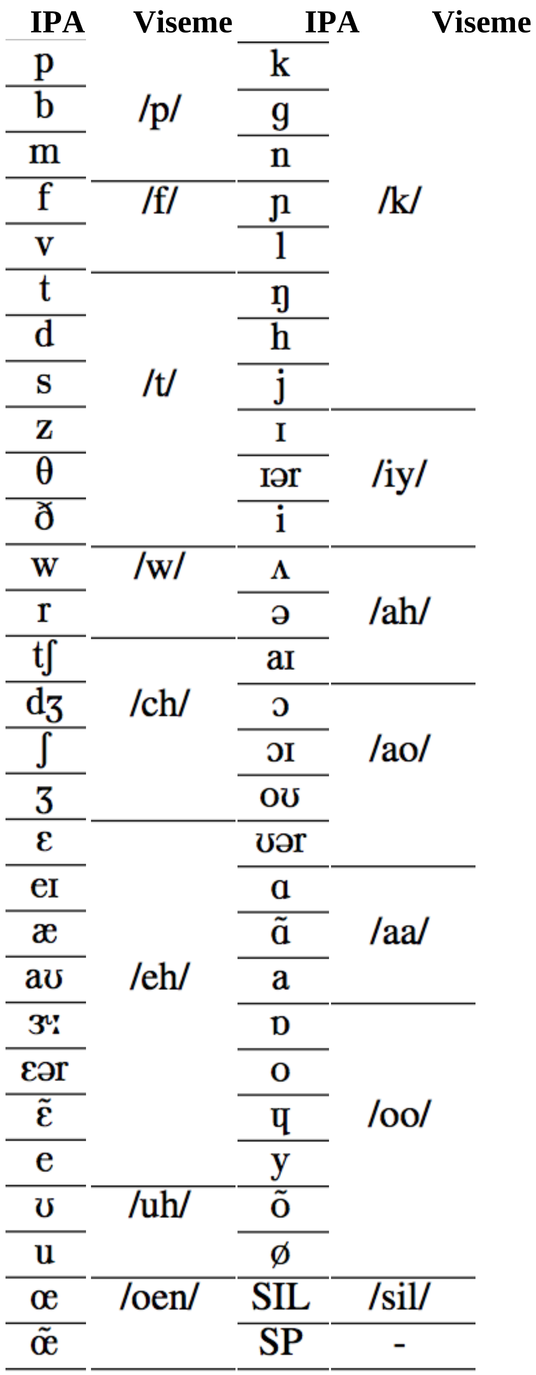
All speech information including the smallest unit of language (phonemes) to more complex stimuli (words) can be identified visually (Bernstein & Liebenthal, 2014). However, the ability to identify speech from visual information alone (lip-reading) varies substantially across individuals (Bernstein & Liebenthal, 2014). In quiet conditions, both listeners with NH and hearing impaired listeners are relatively successful in identifying speech based on the acoustic signal alone (Gatehouse & Gordon, 1990; Shannon et al., 1995). In contrast, the degree to which individuals are able to identify speech when presented solely with visual information (lip-reading) varies substantially (Bernstein & Liebenthal, 2014). Phonemes which sound similar can be differentiated in the auditory modality due to differences in voicing, place, and manner of articulation. Peelle and Sommers (2015) highlight the example of the syllables /ba:/ and /pa:/ which are both plosive (manner of articulation) and bilabial (place of articulation). /ba:/ is voiced whereas /pa:/ is voiceless meaning these syllables sound sufficiently different, however they share the same viseme meaning that visually they appear very similar. A shortfall of visual information is that it cannot convey important information such as the vibration of the vocal cords and is only able to provide information about place of articulation (Amano & Sekiyama, 1998; Summerfield, 1992). Consequently, the accuracy rates reported for lip-reading in adults with NH are low and vary across individuals and studies. It is expected that people with hearing impairments would need to rely more on visual information and therefore be more proficient at lip-reading than people with NH (Pimperton, Ralph-lewis & MacSweeney, 2017), however for individuals with complete hearing loss, lip-reading alone is not sufficient for speech perception (Summerfield, 1992). Therefore, research has investigated how well people can lip-read different segments of speech. The following Sections summarise people's ability to lip-read phonemes, words, and sentences.

1.8.1.1 Accuracy of lip-reading phonemes and visemes

A viseme is the visual representation (mouth movement) of a group of phonemes, a single viseme can signify multiple phonemes as shown in Table 2.1 (Bear & Harvey, 2017). Phonemes which share the same viseme such as 'B' and 'P' are difficult to distinguish when lip-reading as they look the same, whereas different phonemes such as; 'T' and 'P' sound similar but can be distinguished by their different visemes. This observation demonstrates how visual information can be both complementary and redundant in relation to auditory information.

Table 1.1

IPA Phoneme to viseme mapping, adapted from (Newman & Cox, 2012)



IPA	Viseme	IPA	Viseme
p	/p/	k	/k/
b		g	
m		n	
f	/f/	ɲ	
v		l	
t	/t/	ŋ	
d		h	
s		j	
z		ɪ	/iy/
θ		ɪər	
ð		i	
w	/w/	ʌ	/ah/
r		ə	
tʃ	/ch/	aɪ	
dʒ		ɔ	/ao/
ʃ		ɔɪ	
ʒ		oʊ	
ɛ	/eh/	ʊər	
eɪ		ɑ	/aa/
æ		ɑ̃	
aʊ		a	
ɝː		ɒ	/oo/
ɛər		o	
ɛ̃		ɥ	
e		y	
ʊ	/uh/	õ	
u		ø	
œ	/oen/	SIL	/sil/
œ̃		SP	-

Individuals with congenital hearing loss would be expected to have good lip-reading skills due to a reliance on visual information. However, Low lip-reading performance has been reported for individuals with congenital hearing loss as accuracy for identifying phonemes in nonsense syllables was 21-43% (Auer, Bernstein, Waldstein & Tucker, 1997) and 19-46 % (Owens & Blazek, 1985). However, high performance has been observed for some individuals with hearing impairments (HI). Benguerel and Pichora-Fuller (1982) compared participants with NH and those with HI on a lip-reading task which required identification of nonsense syllables consisting of vowel, consonant, vowel (VCV). The HI group scored between 67-97% and the NH group scored between 71-99%, however there were no statistically significant differences between the two groups. Walden, Erdman, Montgomery, Schwartz and Prosek (1981) provided seven hours of training to listeners with HI designed to aid them to distinguish between visemes, an improvement of 58% was found after training for identifying consonants. Overall, this suggests that there is variability in lip-reading performance for people with HI; their performance can be less than or equal to that of the NH participants and may improve with training.

1.8.1.2 Accuracy of lip-reading words and sentences

Demorest and Bernstein (1992) presented sentences with visual information only to over 100 participants with NH. They found that the participants identified between zero and 45% of words in sentences correctly. Bernstein, Demorest and Tucker (2000) compared the ability of adults with NH and adults with HI to identify words and sentences on a similar task. Whilst adults with HI outperformed adults with NH on both stimulus types, performance varied from zero to 85% for words correctly identified in sentences. These findings suggest that lip-reading can vary substantially across individuals with NH and individuals with HI. Bernstein et al. (2000) also report that there was only a small subset of adults with HI who were proficient lip-readers, with accuracy ranging from 73% to 88% of correctly identified phonemes in sentences. For the majority of individuals, lip-reading alone was not sufficient to accurately identify phonemes. Auer and Bernstein (2007) built on these findings by testing a larger sample size and it was found that the HI group identified 43% of words in sentences compared to the NH group who correctly identified 18% of words in sentences. Performance on lip-reading tasks can also vary according to the complexity of the stimuli. Stacey et al. (2016) used IEEE sentences which are semantically and syntactically complex, and difficult to predict from context, they found that listeners with NH were only able to identify 2.85% of words in sentences correctly. Overall, these studies suggest that for longer speech stimuli (sentences) some individuals with HI outperform listeners with NH. Low lip-reading performance for people with NH can also be observed but depends on the complexity of the speech stimuli used.

1.8.1.3 Individual differences in silent lip-reading ability

As lip-reading ability in adults with HI can be less than ideal, research has made an effort to design training to support speech perception. However, the results have been successful to varying degrees with some studies showing that individuals who receive training have the equivalent performance of people without training (Summerfield, 1992), whereas, Dodd, Plant and Gregory (1989) found that training improved performance by 10%.

One reason for the variability in lip-reading in individuals with HI is the amount of visual experience they have. Pimperton et al. (2017) compared cochlear-implant users and individuals with NH and found that cochlear-implant users were able to correctly identify more words than NH adults based on the visual information alone. The authors suggest that this was predicted by the onset of deafness coupled with when implantation occurred, as early onset deafness would mean more experience with visual information which could result in enhanced lip-reading ability. Auer and Bernstein's (2007) findings support this as individuals with early onset hearing loss performed better than individuals with NH. However, variability in lip-reading ability was still observed as sentence identification ranged from zero to 85% (Auer & Bernstein, 2007), this suggests that there are other factors which contribute to lip-reading ability.

Several cognitive abilities have been identified which have a relationship with lip-reading such as working memory and processing speed for verbal and spatial information (Feld & Sommers, 2009). A comparison between young and older adults with NH found that younger adults were more accurate at lip-reading sentences which coincided with faster processing speed and better working memory. Any advantage of lip-reading ability in hearing impaired adults seems to decline with age as older adults with hearing impairments were no better at lip-reading compared to age matched individuals with NH (Tye-Murray Sommers & Spehar (2007a). However, hearing impaired adults were better at identifying visual only words meaning that there may be differences in performance depending on the type of stimulus used. Age related differences in lip-reading have also been identified in adults with NH as Tye-murray, Sommers and Spehar (2007b) found that younger adults were more adept at lip-reading consonants, words and sentences compared to older adults.

There has been conflicting evidence regarding gender differences in lip-reading as previous research (Dancer et al., 1994) suggests that females may be better lip-readers than males however, several studies (Auer & Bernstein, 2007; Tye-murray et al., 2007b) did not find support for this claim and suggest that any gender differences that are present may be trivial (Tye-murray et al., 2007b).

How visual information is used to aid speech perception is not fully understood in NH listeners. Understanding the benefits of visual information is particularly important for people with hearing impairments as listening in noise is challenging.

1.9. Visual speech benefit

1.9.1 Visual speech benefit in quiet

Several studies have shown that even in quiet listening situations there is a performance advantage when people are presented with congruent auditory and visual information compared with when they just have access to unimodal speech. Reisberg, McLean and Goldfield (1987) conducted several experiments in which the auditory signal was intact but the content of speech was difficult to understand as it was in a foreign language or in a language they were not fluent in. In this research Canadian participants were asked to listen to and repeat French and German sentences. Performance increased by 15% when the participants were able to see the talker’s face. Arnold and Hill (2001) aimed to replicate Reisberg et al.’s (1987) results and presented French passages to English speakers. They also presented complex passages of speech and asked the participants to complete a comprehension task. Performance improved in both experiments when the participants could see the talker’s face compared to when they could only hear the voice of the talker. This suggests that there is a benefit of visual information when language comprehension is more demanding. In this context, visual information helps to distinguish auditory speech through providing complementary information such as identification of phenomes. Conversely, Jesse, Vrignaud, Cohen and Massaro (2000) found that visual speech information did not improve the ability to interpret from one language to another. Whilst more sentences were correctly identified in the AV condition compared to the auditory only condition there was no significant benefit of visual information in interpreting the speech as performance overall was quite high. The authors suggest that an advantage of visual information would have become apparent if the task difficulty was increased, for example, if sentences were also presented in noise (Jesse et al., 2000).

Visual speech benefit has also been observed with incongruent AV speech and differs depending on the native language of the individual. Massaro, Cohen, Gesi & Heredia (1993) presented the syllables /ba:/ and /da:/ in either an auditory only or an AV condition to speakers of different languages (English, Spanish & Japanese). In the AV condition, the voice onset time of the initial plosive of the auditory syllable was varied to produce either a /da:/ response or a /ba:/ response. It was found that all the participants regardless of language, were better at identifying syllables when presented with AV information compared to auditory only information. These studies provide evidence that visual speech information is important in quiet listening conditions and suggests that the information provided by the visual signal is not redundant.

1.9.2 Visual speech benefit in noise

1.9.2.1 Words in noise

Being able to see a talker’s face may be of most benefit in situations where the auditory signal is degraded. De la Vaux and Massaro (2004) presented single syllable words varying in different levels of completeness in auditory noise. More words were correctly identified in the auditory only condition compared to the visual only condition and, more words were correctly identified in the AV condition compared to the visual only or auditory conditions. Sumby and Pollack (1954) presented words containing two syllables in different SNRs ranging from -30 dB to 0 dB (in steps of 6 dB) in two conditions, auditory only or AV. The number of words correctly identified was similar across both conditions in high SNRs but in low SNRs (high levels of noise) performance was superior in the AV condition compared to the auditory only condition. Indeed, at the most adverse SNRs, performance improved from around 0% correct with auditory only information to 70-80% with AV information. This suggests that in high levels of auditory noise, visual information provides a significant advantage. Other studies have found that AV integration is optimal at mid-levels or low-levels of auditory noise, these studies refer to the increased speech intelligibility on AV trials compared to auditory only trials as visual enhancement. Ma, Zhou, Ross, Foxe and Parra (2009) presented AV words in different SNRs ranging from 0 dB to -25. Visual enhancement was apparent at SNRs of -8 dB to 0 dB. Using a similar paradigm, Ross, Saint-Amour, Leavitt, Javitt and Foxe (2007) found that visual enhancement peaked at -12dB. These studies suggest that there is an optimum level of auditory noise in which visual information confers an advantage over auditory information alone.

Whilst visual enhancement has been demonstrated consistently, the exact visual information used to aid speech perception is unclear. To address this, Jaekl, Pesquita, Alsius, Munhall and Soto-Faraco (2015) presented words in pink noise from 0 to -12 dB (in steps of 3 dB). Words were either presented in an auditory only condition or accompanied by light displays of human faces. Light displays were either isoluminant or luminance contrast, this type of manipulation isolates the dynamic configural information from the face. More words were correctly identified with the addition of luminance contrast light displays. This demonstrates that dynamic facial cues are important for understanding speech in noise and that only a crude representation of the face is needed to enhance speech perception.

1.9.2.2 Sentences in noise

Studies have shown that identifying sentences in noise improves with the addition of visual information. A common way to measure visual speech benefit in noisy listening situations is by determining the SNR at which the participants can identify 50% of the target speech correctly, this is termed the Speech Reception Threshold (SRT). The amount of visual benefit received is defined as the difference in SRTs in dB between the AV and auditory only conditions. For example, if a participant has an SRT of -5dB for auditory only speech but -15dB for AV speech, their visual benefit would be 10dB. MacLeod and Summerfield (1987) presented visual only sentences, and sentences in white noise in auditory only and AV conditions. Visual speech benefit varied across the participants from 6 to 15 dB indicating that speech perception improved with the addition of visual information but there were individual differences in performance. It was also found that there was a relationship between silent lip-reading scores and visual benefit. In a similar study, Grant and Seitz (2000) found that AV sentences were successfully identified more frequently than auditory only sentences. The difference in performance between conditions equates to an increase of 1.6 dB in the AV condition. Whilst small, this is still an improvement. The researchers suggest that visual benefit occurs as movements of the face are synchronised to auditory cues.

In a similar study to Jaekl et al. (2015), Rosenblum, Johnson and Saldaña (1996) used point light displays of faces whilst sentences were presented in noise. Ten SNRs of white noise were used which increased in steps of 3dB and ranged from -27 to 0 dB. The amount of visual information was manipulated as the participants were provided with three different conditions: 1) a full light display of the face 2) lips, teeth and tongue 3) lips only. It was found that the participants were better at identifying speech in noise with the addition of light displays compared to the auditory only condition, and that performance increased as the amount of lights increased. This suggests that even with minimal visual information available (lips only) this is enough to enhance speech perception, and speech perception improves the more visual information is provided.

Overall, these studies suggest that auditory and visual speech information is complementary, meaning information from either modality is not redundant (de la Vaux & Massaro., 2004). As discussed in Section 1.8, some pairs of visemes may look visually very similar and so are difficult to differentiate by sight alone, whereas the same syllables may be easy to distinguish from the auditory signal alone and vice versa. Therefore, when the equivalent visemes and phonemes are presented in unison the visual information aids in helping to distinguish similar sounding phonemes, which may be more useful in noise. For longer speech stimuli such as during conversation in noise, visual speech could be beneficial as it provides cues to segmenting words, stress patterns and prosody (Grant & Seitz, 2000).

1.10 What visual information is used to support auditory speech processing? Evidence from eye-tracking

So far this chapter has outlined how the addition of visual speech information can enhance speech perception. What is unclear is what specifically about visual information influences speech perception and how this visual information is obtained. Eye-tracking can be used to monitor eye movements and thus understand what part of the visual information is important for speech perception. This is integral for a more complete understanding of AV integration (Everdell, Marsh, Yurick, Munhall & Paré, 2007). The high temporal resolution and sampling rates afforded by eye-tracking make it a useful tool for elucidating where people look on a face in real time during speech perception. Eye movements are generally considered a measure of attention, as gaze is almost always focused on the visual stimulus being attended to (Findlay & Gilchrist, 2003) and gaze focuses on stimuli which are relevant for completing a task (Hayhoe & Ballard, 2005). Pupil dilation has also been used as a measure of listening effort during speech in noise tasks. For example, one study found that pupil size increased as auditory noise increased, indicating that the pupil dilates when it is harder to understand speech (Zekveld, Kramer & Festen, 2011).

1.10.1 Evidence that eye-tracking can give important insights into AV speech perception

There are several types of eye-movement measures which are of interest in the current thesis. The first are fixations which can include amount of fixations in a certain area of interest and, fixation duration which indicates how long part of a stimulus was fixated on. Fixations to a particular area are thought to indicate attention to stimuli relevant for completing a particular task (Hayhoe & Ballard, 2005). However, stimuli can also be attended to in peripheral vision (Hoffman & Subramaniam, 1995), so fixations cannot always account for stimuli being processed. The second measure of interest are saccades. Saccades are fast eye movements which can either be planned, for example looking from one location to another, or automatic meaning an eye movement to a novel stimulus. Saccades are also thought to represent changes in attention, Hoffman and Subramaniam (1995) examined how eye movements influence target detection and found that when targets (rectangles) were presented randomly in one of the four corners of the screen, making a saccade to the location of the target increased successful target detection compared to when targets were attended in peripheral vision. Gaze patterns which include multiple saccades can also be assessed to understand how global visual information is used.

Eye movements can provide information about how and when visual information is used to aid speech perception. Mitterer and Reinisch (2017) manipulated visual attention load as videos of talkers uttering incongruent or congruent AV words were displayed in the centre of a screen surrounded by an array of static pictures. The task was to identify the picture which matched the the spoken word uttered in the video. Eye tracking showed that when visual attentional load was increased, the participants looked less at the talker. When visual load was consistent across trials visual cues were used faster, this was reflected in a saccade to the target picture before the onset of the auditory speech. This study provides evidence that the use of visual information in speech perception can be influenced by attentional load.

Cognitive load has also been found to influence where participants look on a face. The participants were presented with incongruent stimuli and asked to report what they heard. In conjunction with this, they had to remember a sequence of numbers, designed to increase cognitive load (Buchan & Munhall, 2012). It was found that when cognitive load is increased during a speech perception task, the McGurk effect occurred less frequently and the participants spent longer looking at the eyes of the talker and less time looking at the mouth.

Different facial regions may be more relevant depending on the visual information needed to complete a specific task. In Lansing and McConkie’s (1999) study the participants viewed visual only sentences and were then required to make judgements about different speech cues including prosody and word segmentation. Movement of the face was manipulated across different facial regions. It was found that different parts of the face were important depending on the task. The upper region of the face was fixated on more often than the lower region of the face and was more useful in providing cues about intonation.

Using the McGurk effect in conjunction with eye movements can shed light on what part of the visual information is important for AV integration. Gurler et al. (2015) divided the participants into strong and weak perceivers of the McGurk effect; strong perceivers experienced the illusion on 50% or trials or more, weak perceivers less than 50% of trials. They found that strong perceivers of the McGurk effect spent longer fixating on the mouth than weak perceivers. Moreover, there was a correlation between the frequency of the McGurk effect and time spent looking at the mouth (Gurler et al., 2015), suggesting that looking at the mouth of a talker is important for AV integration.

1.10.2 Evidence that eye movements may *not* be associated with AV speech perception

Some evidence suggests that there is no relationship between where people look and correct identification of speech. Everdell et al. (2007) were interested in the specific visual information used when gazing at the face of a talker. The study measured the participants' fixations on dynamic faces when sentences were uttered. Seeing a talker's moving face improved speech perception compared to viewing static images. However, there was no correlation between where the participants looked and accuracy of speech perception. The authors suggest that this is evidence that where people look on a face does not influence AV integration. Everdell et al. (2007) additionally found that the participants also had a bias for fixating the right side of a talker's face. This finding has also been found in other studies in which viewers fixate on the right eye more than the left eye (Paré, Richler, ten Hove, & Munhall, 2003; Vatikiotis-Bateson, Eigsti, Yano & Munhall, 1998). A preference for viewing the right side of a dynamic face can be explained by the observation that the upper right side of the face tends to exhibit more movement compared to the left side (Richardson, Bowers, Bauer, Heilman & Leonard, 2000). This means that it may be more informative to fixate the right side of the face during speech perception.

There is some evidence that where people look when presented with incongruent stimuli does not predict perception of the McGurk effect. Paré et al. (2003) conducted several experiments, the first of which found that the participants tended to fixate mostly on the eyes and mouth but that there was no relationship between gaze on these areas and the McGurk effect. In two other experiments the participants' gaze was directed away from the mouth to establish how much influence looking at the mouth has on AV integration. It was found that the McGurk effect only reduced when gaze was directed up to 20 degrees away from the mouth of the talker. This suggests that when looking away from the mouth, access to rich visual information is reduced resulting in a decline in McGurk perception, although this loss of rich information is not necessarily detrimental to speech perception. The McGurk effect was still perceived during this condition, this suggests that sufficient information can be gathered from other areas of the face or from indirect peripheral fixations of the mouth.

1.10.3 Eye movements and speech perception in auditory noise

Findings that eye movements are not related to performance on a speech identification task (Everdell et al., 2007) could be because speech was presented in quiet listening conditions, and where people look on a face may be more relevant in noisy listening situations. Buchan, Paré and Munhall (2008) presented sentences in quiet and in multi-talker babble whilst the participants had to identify key words. Talker identity was also manipulated, in one condition the same talker was used and in another, talker identity changed across trials. In quiet, the participants focused on the mouth and eyes, whereas the number of fixations on these areas decreased in noise. Gaze duration on the nose and mouth increased in noise suggesting that a central location on the face is preferable in noise. It was also found that fixations on the mouth increased when talker identity changed. This highlights the role of gaze in speech perception which is not only to extract visual information to understand speech but also to aid talker identification.

An integral part of communication is the ability to detect emotion, this means that the listener may use gaze for identifying emotion as well as speech. Buchan, Paré and Munhall (2007) compared eye movements during an emotion recognition task and a speech identification task. Both tasks were completed in quiet and with the addition of multi-talker babble. A comparison of fixations across all conditions showed that fixations on the eyes increased in the emotion task compared to the speech task. In noise, the participants tended to look more centrally and fixated on the nose compared to the no noise conditions. There were no differences in time spent looking at the mouth between noise and no noise. This suggests that gaze strategies differ depending on whether visual information is being used for understanding speech or identifying emotions. This study also provides further evidence that fixating centrally on a face may be more informative in noise compared to quiet.

The duration of stimuli used could also influence where people look on a face. The aforementioned studies used sentences in noise whereas Vatikiotis-Bateson et al. (1998) aimed to establish how eye movements change over longer stimuli. Monologues were presented in quiet and three levels of increasing noise which included talkers from different languages and music. Speech intelligibility was measured via multiple choice questions. Overall, the eyes of the talker were fixated on the most compared to other facial features however, fixation duration increased on the mouth as noise increased. One explanation for this is that speech information is not limited to the mouth and encompasses the whole face. The finding that the participants looked at the eyes the most could also suggest that they were trying to glean emotional information from the face (Buchan et al., 2007).

A similar study to that of Paré et al. (2003) manipulated the distance of gaze up to 15 degrees away from the talker's mouth and in a separate condition allowed free viewing of the face (Yi, Wong & Eizeman, (2013). This study used the addition of auditory noise as sentences were presented in quiet and three different SNRs. When gaze was directed away from the mouth speech intelligibility only decreased at a distance of 15 degrees. In the free viewing condition the participants fixated close to the centre of the talker's mouth more often in high noise compared to quiet. However, where the participants looked did not influence speech intelligibility.

In summary, where people look on a face is dependent on the task, whether visual information is being used to identify speech, talker identity or talker emotion. The length of stimuli, and the listening environment either quiet or in noise also influences gaze strategies.

1.10.4 Speech perception with degraded visual stimuli

1.10.4.1 Behavioural findings

To investigate the importance of visual information further, the following research has systematically altered the quality of the visual information to see how this influences speech perception. Brooke and Templeton (1990) degraded videos of a talker's mouth uttering vowels by decreasing the amount of pixels available. When the resolution was less than 32 X 32 pixels the amount of vowels correctly identified reduced. Campbell and Massaro (1997) degraded videos of visemes using spatial quantisation and found that lip reading was still possible despite reduced visual information. Rather than manipulating the image of the talker, Jordan and Sergeant (2000) increased the viewing distance between the viewer and the talker from one to 30 meters and found that accurate speech perception was still possible from a viewing distance of 20 meters. This suggests that visual information from faces can still be influential even when impaired by large viewing distances. An earlier study by Neely (1956) varied the viewing angle and distance between the viewer and talker. This study showed that speech intelligibility was preserved at a distance of nine meters away and the addition of visual information improved performance when speech was presented in noise. Speech intelligibility was higher when viewing the talker head on compared to from an angle or from the side suggesting that access to important visual information is inhibited unless viewed head on. Wozniak and Jackson (1979) manipulated viewing angles of videos of talkers from 0 to 90 degrees and found that this had no influence on the amount of phonemes correctly identified. This study suggests that viewing a face in profile provides enough visual information necessary for accurate speech identification.

Munhall, Kroos, Jozan, and Vatikiotis-Bateson (2004) found no effect of viewing distances when AV videos of sentences were presented up to 3m away. The study also degraded the visual information by manipulating the amount of spatial frequency information available and the auditory information was also presented in mutli-talker babble. It was found that speech intelligibility was enhanced by the addition of the face for all levels of visual degradation except for the most severe, which has the appearance of a line drawing of the face. The clear undistorted face resulted in the most number of words correctly identified in sentences. Tye-Murray, Spehar, Myerson, Hale and Sommers (2016) degraded the auditory signal with multi-talker babble and blurred the visual signal. They found that a degraded visual signal reduced performance on a task in which the participants had to identify target words to complete sentences. McGettigan et al. (2012) used noise vocoding and Gaussian blurring to degrade sentences. The task was to identify the final word in a sentence, and it was found that visual enhancement increased as auditory noise increased. Overall, these studies highlight the benefit of visual information when speech is presented in noise, and shows that reduced spatial frequency information is sufficient for speech perception.

Several studies have also applied these techniques to the McGurk effect. MacDonald, Andersen and Bachmann (2000) applied spatial degradation to McGurk videos which has the effect of making faces appear pixelated. Four different levels of pixellation and the clear image were used. In the highest level of pixilation the participants still perceived the McGurk effect. Fixmer and Hawkins (1998) presented incongruent stimuli in two levels of auditory noise using SNRs of 7dB and 4dB. There were also two levels of visual noise created by attaching a sheet of translucent paper (drafting film) over the computer screen used for stimulus presentation. For the highest level of visual noise, grease-proof paper was used in addition to drafting film. McGurk perception decreased as visual noise increased and as auditory noise increased McGurk perception increased. Thomas and Jordan (2002) conducted several experiments in which the visual information from faces was degraded either by inverting the face and/or using Gaussian blurring to distort the faces. The stimuli included AV congruent words and incongruent words which produce the McGurk effect e.g. auditory /bæt/ with visual /væt/. Words were presented in a clear (un-blurred) condition and three levels of visual blur, the second experiment also included white noise. It was found that the more the image was blurred the more accuracy decreased for AV congruent words. McGurk perception increased with increasing auditory noise and decreased with increasing visual blur. Taken together these studies suggest that fine detail in the features of the face are not necessary for AV integration. When either the auditory or visual stimulus are degraded the participants make more use of the most reliable modality.

1.10.4.2 Eye movement studies

Degrading the visual information and monitoring eye movements can help to establish what visual information is attended to and how that influences AV integration. Wilson, Alsius, Paré and Munhall (2016) manipulated the visual portion of McGurk videos by removing high spatial frequency information which refers to detail in the face, this means that the visual information appeared blurry. Seven levels of blurriness where used as well as the original clear image. Videos were presented in an AV condition and a visual only condition. The McGurk effect was reported more often when the visual signal was clear compared to when it was blurry. Clear images of the talker were more important for visual only (VO) trials than AV trials. Eye movements were recorded on the visual only trials and showed that time spent looking at the talker's mouth did not predict silent lip-reading, however, the participants spent longer looking at the mouth as the quality of the visual information increased. Overall, this study shows that the visual benefit gained from seeing a talker's face is still apparent even when the visual information is degraded. Alsius, Wayne, Paré and Munhall (2016) examined individual differences in visual benefit when words and sentences were degraded through multi-talker babble and visual blurring. Accuracy for identifying speech in noise increased as the quality of visual information increased. This is in contrast to previous findings (e.g. Wilson et al., 2016) which showed that high spatial frequency information was not necessary for accurate speech perception. The authors speculate that the discrepancy between findings is due to differences in articulation of talkers used across the different studies (Alsius et al., 2016). The talkers used in Wilson et al's (2016) study were English whereas in Alsius et al's (2016) study they were American, different accents could mean speech is easier to discern for some talkers compared to others. Eye movements showed that the participants also looked more at the mouth and eyes as the quality of visual information increased, this supports the findings of Wilson et al. (2016) and suggests that looking at the mouth is important when visual information is of most benefit.

The influence of degraded visual information may be different depending on the type of speech stimuli. Alsius et al. (2017) claim that visual degradation inhibits the McGurk effect more than the identification of congruent speech (e.g. Jordan & Sergeant, 2000). The authors state that this finding is most likely because the presence of congruent visual speech information (even when degraded) enhances the identification of auditory speech whereas the illusion arising from the McGurk effect relies on clear visual information. However, evidence suggests that the McGurk effect was still perceived even when the face of the talker was severely pixellated (McDonald et al., 2000). Using different types of visual degradation may influence the McGurk effect and speech intelligibility differently, Jordan and Sergeant (2000) manipulated the quality of the visual stimulus by varying the viewing distance between the participant and the talker. Further research is required to compare congruent speech with the McGurk effect and using the same visually degraded stimuli.

In conclusion, speech perception is more accurate when the visual signal is clear compared to when it is degraded. A degraded visual signal decreases the McGurk effect. Some findings are contradictory as high spatial frequency information was not necessary for accurate speech perception (Alsius et al., 2016) however other studies found that performance increased as the quality of the visual information increased (Wilson et al., 2016).Time spent looking at the mouth of a talker increases when the visual information is clear compared to when it is distorted. Findings are contradictory as to whether looking at the mouth is important for speech perception or not.

Chapter 2: Overview of thesis

2.1 Summary

Audio-visual (AV) integration involves the combining of auditory and visual information which is often required for every day face to face communication. AV integration may be more beneficial when speech perception becomes difficult; such as when it is harder to hear the voice of the speaker. When the ability to identify speech in noise is reduced, people with NH improve with the addition of visual information; when they can see the talker’s face (Sumby & Pollack, 1954).

People with hearing impairments, such as cochlear-implant (CI) users, also benefit from visual information and may be more adept at AV integration than people with NH (Rouger et al., 2007). However, individuals differ in their visual speech perception ability. One explanation for this is the differences in how people extract visual information, and specifically where they look on a face. The majority of previous literature on this topic has focussed on degrading the auditory stimulus, although research has seen a shift towards exploring how speech perception is influenced when the visual stimulus is degraded (e.g. Alsius, Wayne, Paré & Munhall, 2016). How visual information is used in noise, and why some individuals benefit more from visual information compared to others remains unclear. Through degrading the visual stimulus, we can gain an understanding of which part of the visual speech information is important for AV integration.

2.2 Thesis aims

The overall goal of the thesis is to understand the benefit of visual information when the auditory signal is degraded through noise and to elucidate how auditory and visual information interact when one or both modalities are degraded. The specific aims were firstly to explore a well known measure of AV integration; the McGurk effect (McGurk & MacDonald, 1976). The McGurk effect demonstrates the influence of visual information on the perception of auditory speech. Despite prolific use of the McGurk effect in multisensory research, the factors which contribute to variability in the frequency of the illusion are unclear and recent evidence has suggested that the McGurk effect may not be a good measure of AV integration. Understanding how different methodological factors influence McGurk perception is important for researchers wishing to study the McGurk effect. The second aim of the thesis is to examine AV integration when speech is degraded both visually and acoustically, to understand which part of the visual speech stimulus is important. Within this, different types of speech and levels of noise would be explored as the type of noise used can affect the information in the speech signal (Peelle, 2018). Eye movements will also be examined as where people look on a face may determine the quality of visual information they receive. This will clarify inconsistency in the literature as to how eye movements relate to AV integration during speech perception. In this thesis the auditory signal is degraded in two ways, 1) by adding white noise, and 2) by using a vocoder which is designed to degrade speech and simulate the information provided by a cochlear-implant. Examining different noise types is relevant for individuals with hearing impairments and cochlear-implant users in particular, and could provide the groundwork for future research which could design training programmes for people with hearing impairments. The third main aim of the thesis is to provide insight into theories of AV integration through investigating the timing of AV integration using reaction time. Whilst the McGurk effect is evidence that visual information influences auditory speech information, the mechanisms behind how visual information influences auditory information are not well understood.

Four experiments are reported which address these aims which can be summarised as follows.

The main aims are to assess:

1) The McGurk effect as a measure of AV integration and to explore how different methodological factors can influence the McGurk effect
2) AV integration and the benefit of visual information in quiet and with degraded auditory and visual stimuli
3) AV integration theory by investigating the timing of AV integration

Chapter 1: Audio-visual integration

This chapter discusses research into AV integration starting with an overview of AV integration with non-speech auditory and visual stimuli. The influence of visual information is reviewed including research using the McGurk effect and congruent speech stimuli. Key theories of AV integration are outlined. Research using eye tracking is then discussed including how eye movements can influence AV integration. Two types of speech perception are discussed 1) Auditory speech perception and the difficulties of listening in noise, and 2) visual speech perception, for both listeners with NH and listeners with hearing impairments.

Chapter 2: Overview of thesis

This chapter provided a brief summary of the thesis, outlines the thesis aims and summarises the contents of each chapter.

Chapter 3: Experiment 1

This chapter outlines Experiment 1, which explored a widely used measure of audio-visual integration; the McGurk effect. Different tasks (forced choice vs. open-set) and stimuli were compared to establish how they influence the frequency of the McGurk effect. It was found that the frequency of the McGurk effect varied according to the task type (forced choice or open-set), different stimuli, participants, and how the McGurk effect was defined. Taken together, these factors could account for different estimates of the McGurk effect in previous research.

Chapter 4: Experiments 2

Chapter four outlines an experiment which used behavioural and eye-tracking methods to understand how AV integration is influenced when both the auditory and visual modalities are degraded. Incongruent stimuli were presented in different levels of auditory noise and visual blurring. Eye movements were recorded because where people fixate on a face may also influence the quality of visual information provided. Experiment 2 degraded the auditory stimuli using white noise.

Chapter 5: Experiment 3

Experiment 3 used the addition of vocoding to simulate noise experienced by a cochlear-implant user. The chapter considers speech perception for listeners with NH and listeners with hearing impairments with a particular focus on cochlear-implant users. Research with cochlear-implant users is compared to the findings of research using vocoded speech with listeners with NH. Experiments 2 and 3 showed that when the visual stimulus was clear, AV integration increased as measured by an increase in the frequency of the McGurk effect. Fixating the mouth was not necessary for AV integration to occur, but AV integration increased when the mouth was fixated compared to when it was not.

Chapter 6: Experiment 4

This chapter outlines Experiment 4 which used another measure related to AV integration; visual benefit, to further understand how visual information is used in degraded listening conditions. Word stimuli were degraded using auditory noise and visual blur. Eye movements were recorded to establish if where people look on a face influences the amount of visual benefit received. Visual benefit increased as the clarity of the visual information increased and as auditory noise increased. RT was faster on AV trials compared to auditory only when the visual stimulus was clear.

Chapter 7: Discussion

The findings of the four experiments are summarised and discussed in the context of the wider literature relating to AV integration using degraded stimuli. The implications of the results are outlined including implications for understanding AV integration, theories of speech perception, individuals with hearing impairments, and methodology which could be used in future experiments. Ideas for future research are presented, including investigating AV integration with different age groups and individuals with hearing impairments.

Chapter 3: Experiment 1

3.1 Introduction

Experiment 1 addresses the first and third aims of the thesis, to explore the methodology associated with a measure of AV integration; the McGurk effect, to gain a better understanding of factors which influence perception of the illusion and to understand the timing of AV integration.

3.1.1 Variability in the McGurk effect

Not everyone perceives the McGurk effect, and despite extensive study, the prevalence of the McGurk effect is difficult to determine. A recent review reported that estimates of the McGurk effect range from 32% to 98% across different studies (Alsius et al., 2017), with the original McGurk and MacDonald paper (1976) reporting 98% of illusory percepts. The term McGurk perception will be used to refer to any instances when individuals perceive an illusory percept. McGurk perception varies substantially across individuals with some consistently perceiving the McGurk effect across trials and others never perceiving the effect at all (Basu-Mallick et al., 2015; Gurler et al., 2015; Nath & Beauchamp, 2012). There are numerous individual differences that could explain this variability which are often beyond the scope of a single study to take into account. Several populations have been identified as experiencing a reduced McGurk effect such as those with psychiatric disorders (e.g. schizophrenia; White et al., 2014), dyslexia (Bastien-Toniazzo, Stroumza & Cavé, 2010) and autism spectrum disorders (ASD; Ujiie Asai, Tanaka, Asakawa & Wakabayashi, 2014). It is unclear if an individual’s auditory or visual experience may also influence susceptibility to the illusion for example, musicians who had 13 years’ experience of playing an instrument did not experience the McGurk effect (Proverbio, Massetti, Rizzi & Zani, 2016). This would suggest that expertise in the auditory modality changed the weighting of the auditory and visual senses so that the visual information did not influence the auditory information sufficiently to produce the illusion. However, a recent replication (Politzer-Ahles & Pan, 2019) refuted this claim and found that musicians experienced the McGurk effect to the same extent as non-musicians.

The abilities to lip read and detect AV incongruence have also been correlated with McGurk perception (Strand et al., 2014). The superior temporal sulcus (STS) has been identified as a brain area important for perceiving the McGurk effect, in particular, individuals who perceived the McGurk effect more often also had greater activation in the left STS compared to individuals who perceived the effect less often (Nath & Beauchamp, 2012). Given the extensive list of factors associated with individual differences in the McGurk effect it is important to identify a methodology which increases the likelihood that people will perceive the illusion and reduces variability. This would help to establish if the McGurk effect is an appropriate measure of AV integration.

Several methodological factors have also been investigated which could account for different rates of prevalence of the McGurk effect. Notably, estimates of the McGurk effect appear to depend on the stimuli used, the experimental procedures employed, and differences between participants (Basu Mallick et al., 2015). Basu Mallick et al.’s (2015) study compared 12 different incongruent stimuli which were used in previous studies and found that McGurk responses ranged from 17-58% across different stimuli. Differences in the properties of the stimuli, such as being recorded by different talkers, account for 50% of the variance in McGurk perception (Jiang & Bernstein, 2011). The stimulus set size can also influence the extent of the McGurk effect for some stimuli but not others, for example; when specific pairs $A_{MA}V_{NA}$ and $A_{PA}V_{KA}$ are presented as part of a small set size (2 incongruent stimuli) the McGurk effect was reported more often compared to the medium (4 incongruent stimuli) or large set size (8 incongruent stimuli; Amano & Sekiyama, 1998).

Differences in the type of task used could also influence McGurk responses, studies either use a forced-choice task (e.g. Alsius, Möttönen, Sams, Soto-Faraco, & Tiippana 2014; Colin et al., 2002; Sekiyama et al., 2014; van Wassenhove, Grant & Poeppel, 2005) or an open-set task (e.g. Nath & Beauchamp, 2012). A forced-choice paradigm is most commonly used with either two (e.g. Brancazio & Miller, 2005), three (e.g. van Wassenhove et al., 2005) or four response options (e.g. Colin et al., 2002). Comparisons of open-set and forced-choice procedures have found that forced-choice tasks result in an increase in McGurk responses compared to open-set tasks (Colin, Radeau & Deltenre, 2005; Massaro, 1998). Basu Mallick et al. (2015) found that a forced-choice task increased fusion responses by 18% compared to an open-set task. A limitation of forced-choice tasks, is that when the participants’ responses are constrained, it could be that they are experiencing an illusory percept other than the defined responses but are unable to express this. Therefore, the advantage of an open-set procedure is that it provides the opportunity for the participants to articulate exactly what they heard.

Asking the participants to report how confident they are in their response allows exploration of what the participants perceive. In a lip-reading task Easton and Basala (1982) presented incongruent AV words and asked the participants to report what they saw. Confidence was assessed with a five-point scale, with higher scores indicating increased confidence in the visual signal. They found that the participants were more confident when they correctly identified the visual word than when they were incorrect. McGurk fusion responses were also reported for some words, for example; V_{mail} with A_{but} resulted in the percept ‘bell’. The participants were more confident when they reported the auditory word compared to when they reported McGurk fusion responses. Amano and Sekiyama (1998) compared confidence ratings on a scale of one to five according to different stimulus set sizes and found that auditory responses, and confidence in responses increased as set size increased. Therefore, using fewer stimuli resulted in increased McGurk responses but less confidence in those responses. These studies suggest that when AV information is incongruent, the participants tend to place confidence in the auditory modality, which may be more informative than the visual modality as it contains information about place, manner and voicing which can help with the identification of consonants (Lisker & Abramson, 1964). Using confidence ratings in conjunction with a forced-choice and open-set task would help to establish what the participants perceive and which modality they find more reliable.

The McGurk effect can also be used to establish at what point auditory and visual information converge, as McGurk stimuli are incongruent and short (~2000ms) this allows the time course to be easily investigated. For example; van Wassenhove et al. (2005) used the McGurk effect in conjunction with EEG and found that AV integration occurred within ~50-100ms after stimulus onset. Reaction time (RT) can also be used to assess the timing of speech processing. Congruent AV speech stimuli resulted in faster RT compared to auditory only stimuli (Sumby & Pollack, 1954) whereas incongruent auditory and visual information resulted in slower RT when the participants performed an object categorisation task (Giard & Peronnet, 1999). Studies have shown that RT is slower for incongruent stimuli compared to congruent AV stimuli (Beauchamp, Nath & Pasalar, 2010; Green & Gerdman, 1995; Sekiyama et al., 2014). Longer RT in response to incongruent stimuli may reflect differences in processing involved for incongruent stimuli compared to congruent speech.

Moris Fernández, Macaluso and Soto-Faraco (2017) found that incongruent stimuli activated areas of the brain which relate to speech conflict such as anterior cingulate cortex and inferior frontal gyrus, these areas were more strongly activated when the McGurk effect was perceived compared to when it was not. This provides support that the longer RT associated with incongruent stimuli may reflect the conflict resolution involved in perception of the McGurk effect. Massaro and Cohen (1983) posited that RTs reflect decision making and that congruent information in both modalities would result in faster RT whereas incongruent information would result in slower RT due to the additional time needed to resolve the inconsistency in each modality. This view is consistent with the fuzzy logic model of perception (FLMP) which states that auditory and visual information is processed separately and integrated later in time to form a unitary percept (Massaro & Cohen, 1983).

Alternatively, longer RTs could be indicative of the individual's ability to detect incongruent AV information as Benoit et al. (2010) found that RTs were longer when incongruent stimuli were judged as incongruent compared to when they were judged as congruent. Longer RTs could also reflect the participants' uncertainty in what was heard - using confidence ratings in conjunction with RT would help to clarify this. Comparing RT depending on whether or not the McGurk effect is perceived could shed light on the temporal processing associated with AV integration.

3.1.2 Aims

A well-known measure of AV integration, the McGurk effect will be explored. [add in 1-2 sentences just reframing why I want to look at McGurk effect] Variability in estimates of the McGurk effect has been reported but the factors which influence perception of the illusion are not well understood. The goal of this experiment was to inform methods by exploring differences in McGurk perception between the participants, stimuli and task type. The aims were: a) to test the same participants on different procedures (open-set vs. forced-choice tasks), to see the influence on McGurk responses; b) to test which stimuli produce the McGurk effect to the greatest extent to inform future experiments in the thesis; c) to examine confidence ratings about what is perceived and d) to assess RT in relation to incongruent and congruent speech. In order to examine the influence of task type (open-set or forced-choice) on McGurk responses, the participants completed both tasks. The order in which the participants completed the tasks was counterbalanced because whether the participants were required to make an open or forced-choice response first could influence their responses in subsequent blocks. Additionally, in order to estimate how reliable the participants' responses were, the participants completed each task type twice and reported their confidence in their responses. This experiment builds on previous research through providing a more detailed analysis of open-set and forced-choice responses. The addition of confidence ratings will also help to assess which task type is most the appropriate. Furthermore, in Basu Mallick et al.'s (2015) study different groups of participants were used for each task type (open-set vs. forced-choice), therefore any differences in McGurk perception could be due to individual differences. The results of this experiment can also be used to inform which stimuli and method to use in subsequent experiments in the thesis.

3.1.3 Hypotheses

It is expected that in line with previous literature a) the forced-choice task will result in increased McGurk perception; b) McGurk perception will vary across talkers with some talkers producing the illusion to greater extents than others; c) different individuals will vary in the extent to which they perceive the McGurk effect; d) the participants will be more confident of their responses on the open-set than the forced-choice blocks as they will have more freedom to choose their response and e) RT will be slower for incongruent stimuli compared to congruent stimuli.

These findings would provide further evidence that variability across individuals and task type influences how often the McGurk effect is reported. This has implications for researchers wishing to use the McGurk effect as a measure of AV integration. Assessing these hypotheses will contribute to our understanding of the methodological factors which influence the reports of AV integration as measured by the McGurk effect.

3.2 Method

3.2.1 Design

This experiment employed a 2 x 2 x 2 mixed design, the within-participants independent variable was Task Type (Open-set or Forced-choice) and the between participants independent variables were block order (Open first or Forced first) and Block Presentation (First, Second). There were three separate dependent variables 1) whether the participants perceived the McGurk effect, this was classified as either (a) fusion responses, or (b) any non-auditory response, 2) RT (ms) and 3) confidence ratings (on a scale of 1-7).

3.2.2 Participants

The participants were 46 students from Nottingham Trent University, 5 males and 41 females aged 18 -35 years (M = 21.30), the sample size was based on opportunity sampling. The project was approved by the Social Sciences Research Ethics Committee. The participants gave informed consent and received course credits for their time. The informal inclusion criteria was that the participants were native English speakers and reported NH, and normal or corrected to normal vision. The informal exclusion criteria included the participants who reported a diagnosis of dyslexia or Autism. This was important, as individuals with dyslexia or ASD have been shown to experience a reduced McGurk effect (Bastien-Toniazzo et al., 2010; Saalasti, Tiippana, Kätsyri & Sams, 2011).

3.2.3 Stimuli & Apparatus

Stimuli consisted of videos of 10 women uttering the syllables: /ba:/, /da:/ and /ga:/. The talkers were aged between 25 and 40. They wore black and were filmed in front of a white background in a quiet room. Materials were recorded using a Panasonic AVC HD video camera, and auditory stimuli were recorded using a Studio series SL150 microphone.

Stimuli were edited using Adobe Premiere Pro version 9.0. A static face of the talker was added to the start and end of each video to increase the overall length of the video. Once edited, each stimulus was ~2000ms in duration. Auditory stimuli were sampled at 41000 Hz with 16-bit quantization, and the video files had a resolution of 720 x 526 pixels. For each talker, 5 stimuli were produced: 3 congruent stimuli which consisted of auditory and visual /ba:/, /da:/, and /ga:/, and 2 exemplars of incongruent stimuli which consisted of auditory /ba:/ and visual /ga:/ ($A_{BA}V_{GA}$). incongruent stimuli were created by dubbing the auditory /ba:/ stimuli onto the visual /ga:/ stimuli as shown in Figure 3.1. The audio track was overlaid over the video so that the auditory utterance appeared synchronised with the visual mouth movement, this was achieved by aligning the acoustic burst of the auditory stimulus with the acoustic burst of the video. The talkers from each stimulus are shown in Appendix A.

All stimuli were presented at the same sound level (average 70dB SPL) determined by using an artificial ear to measure sound levels over headphones (Brüel & Kjær Type 4153). Stimuli were presented via a 17inch computer screen with a resolution of 1920 x1080 pixels and the videos filled 75% of the screen. Stimuli were presented via EPrime (Version 2.0, Psychology Software Tools Inc., Sharpsburg, US) and using HD280pro headphones (Sennheiser, Wedemark, Germany) via a custom built digital-to-analogue converter.

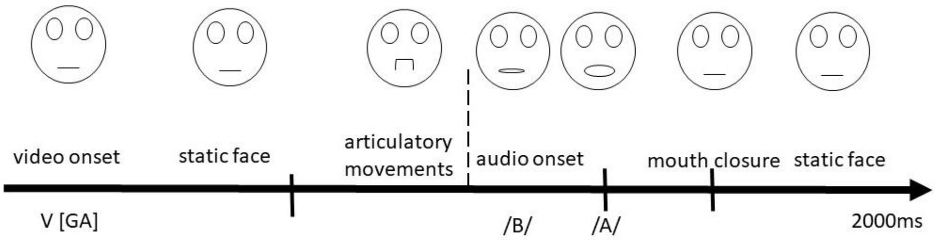


Figure 3.1. Schematic representation of incongruent stimuli showing the visual onset and voice onset (van Wassenhove et al., 2005).

3.2.4 Procedure

The participants sat in front of a desk ~45cm away from the computer. Before the experiment began, the participants were instructed to watch the videos closely, listen carefully and then respond by repeating out loud what they heard (open-set), or by pressing 1 of 3 keys labelled with ‘BA’, ‘GA’, or ‘DA/THA’ (3 option forced choice task); Basu Mallick et al., 2015). Key placement was counterbalanced across the participants and responses were recorded using a Dictaphone. There were 12 practice trials (videos) before the start of each of the experiment. Practice trials consisted of congruent stimuli only which were videos recorded by two of the talkers (3 congruent stimuli, 2 talkers, repeated twice). The other eight talkers were used for the test trials. Each of the eight talkers had 5 stimuli (2 McGurk + 3 congruent), giving 40 trials per block, 8 x 5 = 40 trials in total in each block. These 40 test trials were presented 4 times (160 in total) in alternate blocks e.g. open-set, forced-choice, open-set, forced-choice. After the video appeared the participants were asked to either respond out loud or press a button on the keyboard. Reaction time was recoded for key presses. After this a subsequent screen appeared asking the participants to rate their confidence in what they heard on a scale of one to seven. The condition order was counterbalanced across the participants so that half (*N*=23) completed the forced-choice block first and half (*N*=23) completed the open-set block first. This was to prevent the participants who received a forced-choice block first from being influenced in the open-set block by adapting their responses to fit within the three options specified in the forced-choice block. Data were recorded automatically through the experimental software (E-prime).

3.3 Results

3.3.1 Summary of responses

The average correct responses for congruent stimuli were: /ba:θ/ (M = 94%, SD = 7%), /da:/ (M = 96%, SD = 5%) and /ga:/ (M=91%, SD = 6%). The percentages of each response were averaged across all examples of incongruent stimuli (the two tokens from the 8 talkers) for the open-set task. The participants reported 14 different percepts including: /a:/ (11%), /la:/ (0.50%), /ba:θ/ (0.17%), /gla:/(0.03%), /gɔ:/ (0.06%), /pa:/(0.13%), /ta:/(0.17%), /ɔ:/(0.17%), /bra:/ (0.10%), /bwa:/ (0.03%). The most frequently reported syllables on open-set trials corresponded to the auditory /ba:/ (42%), visual /ga:/ (25%), fusion /da:/ (16%) and /θa:/ (13%), as in the ‘th’ in think. McGurk responses across open- and forced-choice tasks were then coded in two ways, as anything other than the auditory (/ba:/) and fusion responses only (/da:/,/θa:/). The overall mean percentage of non-auditory responses was 57.5% (*SD* = 20.7%) whereas for fusion it was 17.5% (*SD* = 12.3%).

3.3.2 Variation in McGurk responses across stimuli

Figure 3.2 shows the fusion and non-auditory responses for the different stimuli. Different talkers elicited McGurk responses to different extents, and this also depended on the definition used (Talkers 1 and 2 produced the most non-auditory responses, while Talkers 4 and 6 produced the most fusion responses).

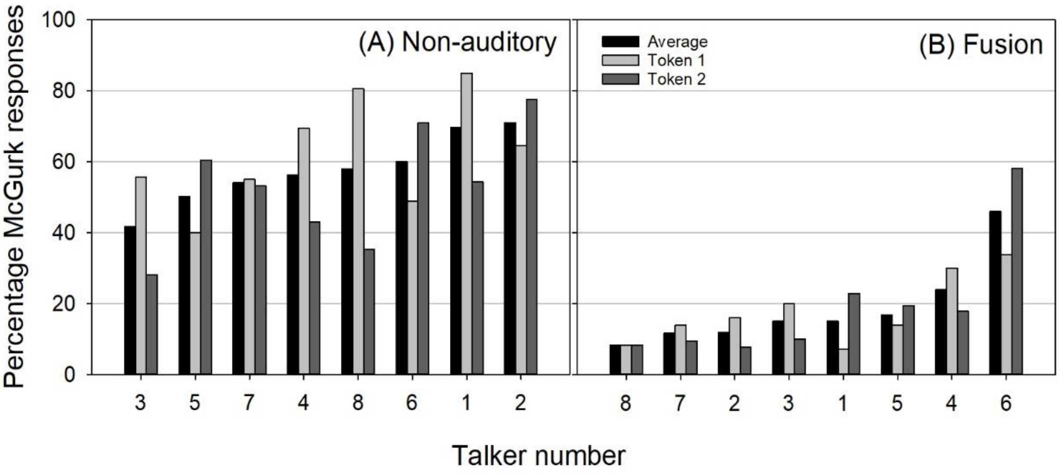


Figure 3.2. Percentage of McGurk responses given by stimuli recorded by different talkers. Stimuli have been ordered by the average number of McGurk responses, from fewest to most. Panel A shows Non-auditory responses, and Panel B shows Fusion responses.

3.3.3 Variation in McGurk responses across participants

Figure 3.3 depicts the percentage of McGurk responses according to fusion and non-auditory responses ranked from smallest to largest. There was a large amount of variation in McGurk responses with some participants not perceiving the effect at all and the strongest perceivers experiencing the illusion on 52% (fusion) and 98% (non-auditory) of trials.

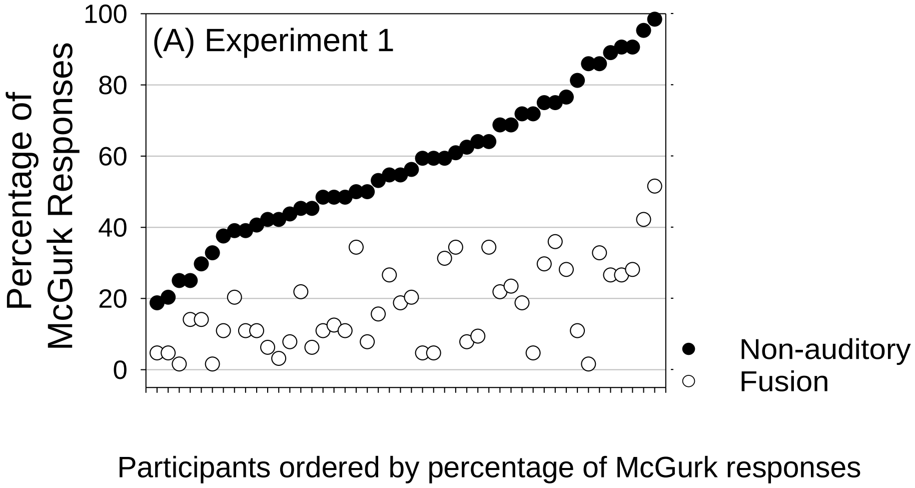


Figure 3.3. Percentage of McGurk (non-auditory and fusion) McGurk responses across the participants. The participants have been ordered by the percentage of McGurk responses they gave, from lowest to highest with each tick representing a separate participant's response.

3.3.4 Differences between open-set and forced-choice responses

The percentage of McGurk responses in Open-set and Forced-choice blocks was calculated and the results are shown in Figure 3.4. Panels A and B show McGurk responses analysed with the Non-auditory definition, while Panels C and D show responses analysed with the Fusion definition. Separate 2 x 2 x 2 mixed ANOVAs were conducted for fusion responses and non-auditory responses with the within-participants independent variables Task Type (Open-set or Forced-choice), block presentation (first or second) and the between-participants independent variable block order (Open first or Forced first). Overall, with the Non-auditory definition of the McGurk effect people made significantly fewer (*M* = 51.7%, *SD*= 22.7%) McGurk responses on the Forced-choice blocks compared with the Open-set blocks (*M* = 61.9%, *SD* = 24.2%; *F*(1, 44)=46.37, *p*< .001, eta squared η^2 = .087). Additionally, block presentation was included as the participants may exhibit learning effects when the stimuli are repeated. Non-auditory McGurk responses increased on the second presentation of the stimuli (*F*(1,44)=8.64, *p*=.005, η^2 = .014). Block order (open first or forced first) was included as a between subjects variable as whether the participants were required to make an open or forced-choice response first may have influenced their responses in subsequent blocks. There was no significant main effect of block order (*F*(1,44)=2.04, *p*=.16, η^2 = .044) and no significant interaction between task type and block order (*F*(1,44)=2.66, *p*=.11, η^2 = .110), no significant interaction between block presentation and block order (*F*(1,44)= .39, *p* =.53, η^2 = .001), task type and block presentation (*F*(1,44)= .07, *p* =.79, η^2 < .001), and no significant interaction between all three task type, block presentation and block order (*F*(1,44)= .07, *p* =.79, η^2 < .001).

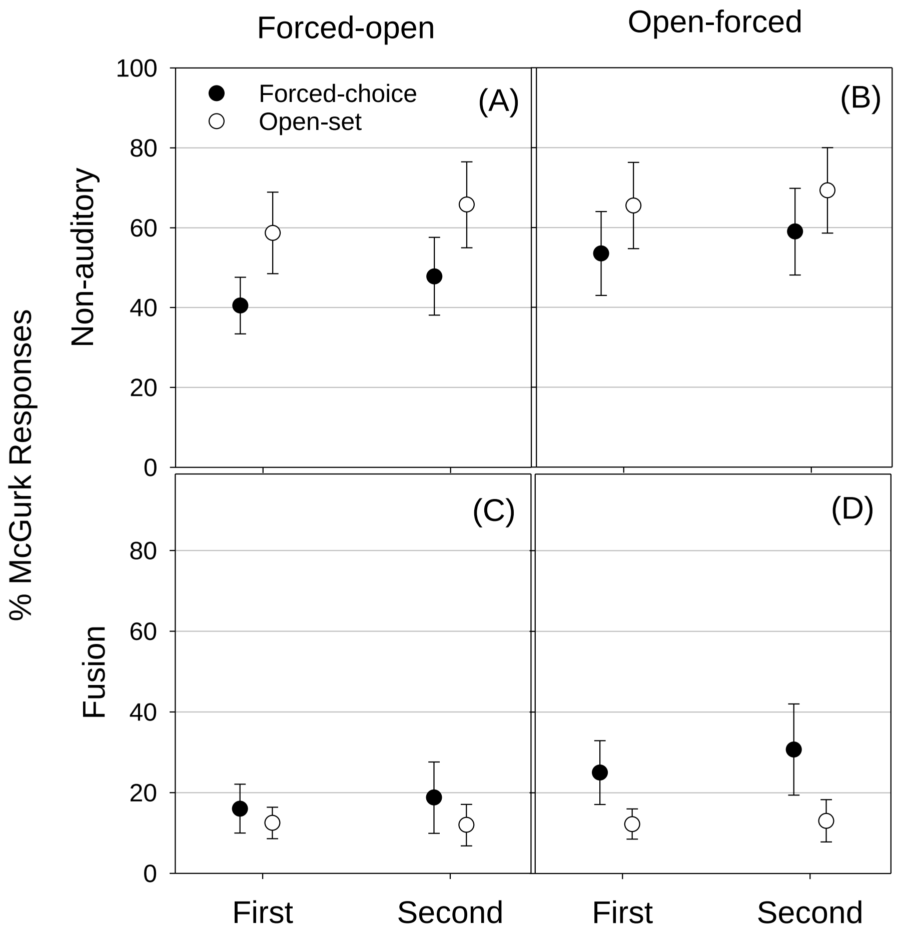


Figure 3.4. McGurk responses according to task type and definition. The same data were coded according to two definitions of McGurk responses; non-auditory and fusion, for Forced-choice and Open-set task types. Panels A and B show McGurk responses classified according to the Non-auditory definition, and Panels C and D show the Fusion responses. Fusion responses are a more conservative estimate of McGurk responses than the non-auditory definition which allows for a broader range of responses. Panels A and C show the Forced First-Open second blocks, and Panels B and D show the Open First-Forced second blocks. Error bars represent 95% confidence intervals.

Using the Fusion definition, the opposite pattern was found, where significantly *more* McGurk responses were made on Forced-choice blocks (*M* = 22.6%, *SD*= 20.7%) than on Open-set blocks (*M*=12.4%, *SD*= 10.4%, *F*(1,44)=17.50, *p*<.001, η^2 =.089).

Overall, fusion responses did not significantly increase on the second presentation of the stimuli (*F*(1,44)=2.59, *p*=.11, η^2 = 0.04) and there was no significant effect of block order (*F*(1,44)=2.31, *p*=.14, η^2 = .050). There was however a significant interaction between Task type and Block order (*F*(1,44)=4.26, *p*=.045,η^2 = .022). As Figure 3.4 shows, there was a larger difference between Forced-choice and Open-choice blocks for the participants with the Open-Forced order (15%) than for the participants in the Forced-Open group (5%). A 2 x 2 repeated measures ANOVA on the Forced-Open group revealed no significant effect of Task type (*F*(1,22)=3.64, *p*=.069, η^2 = .032). However, a 2 x 2 repeated measures ANOVA for the Open-Forced group did reveal a significant effect of Task Type (*F*(1,22)=14.11, *p*=.001, η^2 = .161). Therefore, the effect of Task type was driven by the participants who completed the Open-set task before they completed the Forced-choice task.

3.3.5 Confidence ratings

Confidence ratings were measured on a scale of one to seven, one meaning not at all confident and seven meaning highly confident. Overall confidence ratings for the congruent stimuli /ba:/ (*M*=6.19, *SD*=.71) /ga:/ (*M*=6.24, *SD*=.72) and /da:/ (*M*=6.29, *SD*=.72) were similar indicating that the participants were very confident about their responses but less confident about their responses to incongruent stimuli (*M*=4.71, *SD*=, 1.02). A one-way ANOVA with four levels according to the different stimulus types (BA, GA, DA, $A_{BA}V_{GA}$) showed that there was a significant effect of Stimulus type (*F*(1.42, 64.23)= 146.27, *p* <.001, η^2 = .407). Pairwise comparisons indicated that the participants were more confident when the stimuli were congruent compared to when they were incongruent (*p* <.001) there were no significant differences in confidence ratings between /ba:/ and /ga:/ (*p*=1.00), /ba:/ and /da:/ (*p*=.265), /ga:/ and /da:/ (*p*=1.00). Confidence ratings on incongruent trials (Table 4.1) were compared. A paired samples t-test showed there were no significant differences in confidence ratings when the participants perceived the McGurk effect compared to when they did not *t*(45)=1.72, *p*=.092, *d* = .013.

Confidence ratings for incongruent stimuli were similar across all blocks. For the forced-choice blocks mean confidence was 4.7 on both the first and second presentations (*SD* = 1.8, 1.9 respectively). For the open-set blocks mean confidence ratings for incongruent stimuli were 4.5 (*SD* =1.8) on the first presentation and 4.8 (*SD*=1.7) on the second presentation. A 2 (Task Type: Open or Forced) x 2 (Block Presentation: First or second) x 2 (Block order) mixed ANOVA found no significant effect of Task type on confidence ratings (*F*(1,44)=.119, *p*=.732, $\eta^2 < .001$) and no significant effect of Block presentation (*F*(1,44)= 3.45, *p*=.070, $\eta^2 = .005$) or Block order (*F*(1,44)=.132, *p*=.718, $\eta^2 = .003$).

Table 3.1

Mean confidence ratings for incongruent stimuli with standard deviations

		Block 1	Block 2	Average
Forced-open	Open-set	4.6 (1.7)	4.6 (1.7)	4.6 (1.7)
	Forced-choice	4.5 (1.7)	4.7 (2.0)	4.6 (1.8)
Open-forced	Open-set	4.4 (1.9)	4.9 (1.7)	4.6 (1.8)
	Forced-choice	4.8 (1.8)	4.8 (1.9)	4.8 (1.8)
Average		4.6 (1.8)	4.7 (1.8)	

3.3.6 Reaction time to incongruent stimuli

The participants were slower to respond when the McGurk effect was perceived (*M*= 2663.57ms, *SD*= 629.06) compared to when it was not (*M*= 2627.67ms, *SD*= 477.46) but a paired samples t-test showed this was not significant *t*(45)= -.41, *p*=.679, *d* = .023. RT on congruent and incongruent trials was compared regardless of whether or not the illusion was perceived. Overall RT for the congruent stimuli /ba:/ (*M*=2209.72, *SD*=353.92) /ga:/ (*M*=2013.81, *SD*=301.27) and /da:/ (*M*=2148.07, *SD*=320.68) were faster than responses to incongruent stimuli (*M*=2595.81, *SD*=414.02). A one-way ANOVA showed there was a significant effect of Stimulus type as the participants were slower to respond on incongruent trials compared to congruent trials *F*(2.22,100.01)= 81.46, *p* <.001, $\eta^2 = .275$). Pairwise comparisons showed that responses to all congruent stimulus types (BA, GA & DA) were significantly (*p* <.001) faster than responses to incongruent stimuli.

3.4. Discussion

The aim of this experiment was to clarify how perception of the McGurk effect varies across the different task types, definitions, participants, and stimuli. Confidence ratings were used to explore responses on the different tasks and RT was measured to assess the differences between congruent and incongruent speech.

Overall, it was found that perception of the McGurk effect was highly variable depending on how it was defined, which stimuli were used for testing, and according to which individuals were tested. McGurk responses ranged between 2-52% when using the Fusion definition, and 19-98% when using the Non-auditory definition. The upper bound of fusion responses was much lower than that reported by previous studies, such as the 98% reported by McGurk & McDonald (1976) and the 100% reported by Basu Mallick et al. (2015). This could be in part attributable to differences in the stimuli that were used as different stimuli produced the McGurk effect to different extents, and also depended on the definition used, in future mixed effect models could be used as this type of analysis is able to account for variability across stimuli. Fusion responses occurred between 8 and 43% across different talkers, whilst using the non-auditory definition, the McGurk effect occurred between 42 and 71% across different talkers. Jiang and Bernstein (2011) showed that half of the variance in McGurk perception was accounted for by talker differences, this could include; facial features such as the size of the mouth aperture or articulation of syllables.

McGurk perception also varied substantially across the participants. Whilst the participants reported NH and vision in the present study, individuals might be more inclined to attend to either modality depending on their auditory or visual experience. Those who are more attuned to the auditory modality may have experienced the McGurk effect less (Proverbio et al., 2016).

Consistent with previous studies (e.g. Basu Mallick et al., 2015), it was found that the type of task (open-set or forced-choice) influenced the frequency with which the participants reported perceiving the McGurk effect. When using the Fusion definition, the participants made more McGurk responses (average 10%) in the forced-choice task than in the open-set task. This effect supports Basu Mallick et al. (2015) who also found more fusion responses in a forced-choice task. These results suggest that the forced-choice task constrains responses so that an individual may hear something completely different to the auditory /ba:/ or visual /ga:/ syllable but with the limited response options available in a 3 option forced choice task they are unable to express this and so respond with the only other option ‘da/tha’ thus elevating their fusion responses in the forced-choice task. Consistent with this explanation is the finding that people gave a broad range of different responses in the open-set task and accordingly with the non-auditory definition of the McGurk effect, significantly *fewer* McGurk responses were made in the Forced-choice task than in the Open-set task.

An important feature of the current experiment was that the order in which the participants completed the tasks were counterbalanced as it was expected that completing one type of task could affect responses in the second task. This was supported by the interaction between task type and block order, where it was found that the participants made more fusion responses on Forced-choice blocks if they had already experienced an Open-set response task. It is therefore important to take into consideration previous exposure to incongruent stimuli when assessing the magnitude of the McGurk effect. From the present experiment one cannot determine what is the ‘correct’ definition of the McGurk effect to use, and whether the effect should be defined as anything other than the auditory stimulus, or whether the ‘stricter’ fusion definition should be used. However, what is clear is that it is important to be explicit in all research how the McGurk effect is defined.

Confidence ratings according to open-set and forced-choice tasks were explored to help establish which method is more preferable. Confidence ratings were not significantly different according to task type. This is most likely because the participants tended to choose values in the middle of the scale e.g. 3,4,5. In future, instead of using a scale, the participants could be asked if they feel confident about their response and answer yes or no. If the participants had been more confident in either the open-set or forced-choice task this would have lent support for using that particular task in future experiments.

The participants were more confident in their responses when stimuli were congruent compared to when they were incongruent regardless of accuracy. There are several reasons why the participants may have reported low confidence in incongruent stimuli. Firstly, there is an expectation that speech is congruent (Rao & Ballard, 1999) as this is what is experienced in everyday conversation, and perceiving the illusion may result in uncertainty about what was heard. Secondly, low confidence could be related to the stimulus set-size, Amano and Sekiyama (1998) found that larger set sizes resulted in increased confidence in responses. They define a small set-size as containing two types of incongruent syllables and two types of congruent syllables, whereas the large set-size contained eight types of congruent syllables and eight types of incongruent syllables. In the present study, a small set size (1 type of incongruent syllable and 3 types of congruent syllables) was used. Finally, it could be due to individuals' ability to detect incongruence in the stimuli, for example if an individual is good at detecting when stimuli are incongruent they may be less confident about their response as they are aware that the auditory and visual information are different. This could be explored in future experiments by including additional task instructions which ask the participants to rate the stimuli in terms of how congruent they are.

RT was measured to understand how different stimulus types (congruent vs. incongruent) influence the temporal processing of speech. As expected, RT was slower for incongruent stimuli compared to congruent stimuli in line with previous research (Beauchamp et al., 2010; Green & Gerdman, 1995; Sekiyama et al., 2014). This may be indicative of the extra decision making processes involved with incongruent stimuli (Moris Fernández et al., 2017; Massaro & Cohen, 1983). If this is the case, the results would also speak to the FLMP whereby the auditory and visual streams are processed separately, although this is difficult to determine without the use of EEG measures. Future research could use RT in conjunction with EEG to establish the timing of AV integration. An alternative explanation is that as confidence ratings were lower for incongruent stimuli, the longer RT may reflect uncertainty in the participants' responses. There were no differences in RT according to whether or not the McGurk effect was perceived on incongruent trials, this suggests that RT is longer for incongruent stimuli due to the conflicting auditory and visual information rather than whether or not an illusion is perceived.

3.4.1 Choice of procedure in following experiments

One concern with using a forced-choice task was that McGurk responses might be elevated by forcing the participants to report something other than what they perceived (if the response options do not include their percept). This concern was partly supported by the finding that the participants reported more fusion responses in the forced-choice task than the open-set task. However, this pattern was reversed when the Non-auditory definition of the McGurk effect was used; here there were fewer McGurk responses with a forced-choice task than with an open-set task. Therefore, when a non-auditory definition of the McGurk effect is used it seems that McGurk responses are not artificially inflated by a forced-choice task due to uncertainty. This is supported by the confidence ratings; the participants were no less certain in the forced-choice task than in the open-set task.

Given these findings, and that the thesis aims to access the timing of AV integration using RT, the remainder of the experiments reported in this thesis will be based on forced-choice tasks using the non-auditory definition of the McGurk effect.

3.4.2 Conclusion

In order to move forward, a general consensus needs to be reached for defining the McGurk effect. To encompass all types of incongruent stimuli a definition could be used which classifies McGurk responses as a change in auditory perception which produces a syllable different to that of the voice (Tiippana, 2014). Future research should make every effort to take into account factors which can influence the McGurk effect to reduce variability and consider how their stimuli and task type will influence results.

Chapter 4: Experiment 2

4.1 Introduction

This chapter describes Experiment 2 which evaluated the importance of visual information when both the auditory and the visual stimulus are degraded. McGurk perception and eye movements were examined. Gaussian blurring was used to degrade the face of the talker by reducing the spatial frequency of the visual information, and white noise was used to mask the voice of the talker.

A robust finding is that there is a benefit of seeing a talker's face when understanding speech in quiet. Studies show that sentences and passages of speech were identified correctly more often when presented with the face of the talker compared to the voice only (Arnold & Hill, 2001; Reisberg et al., 1987). However, we are often confronted with noise in our every day environment, for example, trying to understand someone in a noisy coffee shop can be difficult as the reliability of the acoustic information is reduced. In this situation, the visual information provided by the face may be more important compared to quiet listening situations. Studies show there is an advantage of seeing a talker's face (visual enhancement) when speech is presented in auditory noise (de la Vaux & Massaro, 2004; Grant & Seitz, 2000; Jaekl et al., 2015; MacLeod & Summerfield, 1987; Rosenblum et al., 1996; Sumby & Pollack, 1954). When either the auditory of visual modality is degraded this can increase AV integration, this phenomenon is known as The Principle of Inverse Effectiveness (PoIE; Meredith & Stein, 1986). Ma et al. (2009) found that visual enhancement was apparent at SNRs from -8 dB to 0 dB. This is consistent with Ross et al. (2007) who found that visual enhancement peaked at -12dB. This suggests that there is an optimum level of auditory noise at which visual information improves speech perception. Therefore, visual information would be of most benefit when auditory information is degraded by noise. The majority of research has focused on how AV integration is affected when speech is presented in auditory noise. However, it is important to study how AV integration changes when the visual signal is degraded to better understand the benefit of visual information. This is also relevant for understanding how people with visual impairments integrate information. Putzar, Hötting and Röder (2010) found that the participants with visual impairments (cataracts) had reduced AV integration as they perceived the McGurk effect less often than the participants with normal vision with the equivalent lip-reading ability. The present research is also timely due to technological advances with video communication such as Skype where the visual signal is often degraded, which could impact on AV integration and hinder communication especially for older adults or people with hearing impairments who may rely more on visual information.

Several different methods have been used to degrade the visual information, increasing the viewing distance between the talker and listener, and manipulating the angle at which the talker's face is viewed did not influence accurate speech perception (Jordan & Sergeant, 2000; Munhall et al., 2004; Wozniak & Jackson, 1979). Increasing pixilation, which reduces detailed information on the face resulted in fewer instances of the McGurk effect (MacDonald et al., 2000). Reducing high spatial frequency information on the face also reduces the McGurk effect but does not inhibit it completely (Paré et al., 2003; Wilson et al., 2016). Overall, the finding that the McGurk effect was still perceived even in high levels of visual degradation suggests that fine detail on the face is not required for AV integration.

Few studies have degraded both auditory and visual speech information. This is important for understanding which modality is used when one or both are degraded and how this influences AV integration. Munhall et al. (2004) reduced the spatial frequency information on faces and presented auditory speech in mutli-talker babble. Performance was higher for AV speech compared to auditory only conditions except for the highest level of visual degradation. This means that there was still a benefit of seeing a talker's face, even when the quality of the visual information was reduced. In contrast, Tye-Murray, Spehar, Myerson et al. (2016) blurred faces and used multi-talker babble to mask auditory speech. They found that as visual blurring increased performance on a word identification task decreased. Tye-Murray et al. (2010) presented auditory speech in different SNRs and lowered the contrast of the image. Contrary to the PoIE, they found that reducing the quality of information in either modality did not increase AV integration when the participants completed a sentence building task. The conflicting results in these studies may be due to the different tasks used to measure AV integration.

One study presented words which produce the McGurk effect in different levels of visual blurring and white noise (Thomas & Jordan, 2002). AV integration increased as white noise increased, and decreased as visual blur increased. As the McGurk effect is dependent on the visual stimulus, auditory noise may result in more reliance on the visual information which may also increase instances of the illusion.

4.1.2 Eye movements and AV integration

During everyday conversation the listener may attend to either the face or voice of the speaker and this can be dependent on the reliability of the information (Ernst & Bülthoff, 2004; Witten & Knudsen, 2005). In quiet, auditory information is sufficient for understanding speech (Gatehouse & Gordon, 1990; Shannon et al., 1995) whereas visual information can only provide limited speech cues and consequently many individuals find lip-reading difficult (Bernstein & Liebenthal, 2014). Therefore, the most beneficial strategy for the listener is to combine auditory and visual information from the face and voice of the talker, in order to understand speech (AV integration). Where people look on a talking face may be an important factor in explaining variability in AV integration in different situations and across individuals. Gurler et al. (2015) divided the participants into strong and weak perceivers of the McGurk effect. They found that strong perceivers of the McGurk effect spent longer fixating on the mouth than weak perceivers. Moreover, there was a correlation between McGurk effect perception and time spent fixating the mouth (Gurler et al., 2015). In contrast however, Paré et al. (2003) found that fixating the mouth did not predict the extent to which the McGurk effect was experienced. When the participants' gaze was directed 20 degrees away from the mouth the McGurk effect was still present suggesting that fixating the mouth is not always necessary to perceive the McGurk effect (Paré et al., 2003). This finding suggests that face movements which can be seen in peripheral vision are sufficient to produce the McGurk effect.

Gurler et al. (2015) suggested that the contradictory findings may be due to the pre-stimulus fixation cross positioning as their study used a peripheral fixation cross which appeared in one of four corners of the screen whereas Paré et al. (2003) used a central fixation cross. Gurler et al. (2015) argue that the pre-stimulus peripheral fixation cross forces the participants to make a planned eye movement to a particular part of the face whereas a central fixation cross encourages the participants to fixate centrally and attend to other parts of the face in the peripheral vision. Arizpe, Kravitz, Yovel and Baker (2012) used a face recognition task and varied the location of starting fixations when the participants viewed faces. They found that the location of the starting fixation influenced eye movements as saccade latencies were longer when central fixations were used compared to peripheral fixations. These findings suggest that the starting fixation cross used in experiments can influence where people look on a face.

Fixating the mouth and surrounding area may be particularly important when the auditory signal is degraded as this would enable extraction of better quality visual information. When monologues were presented in high levels of background noise including music and multilingual talkers, the participants looked at the eyes approximately half of the time (Vatikiotis-Bateson et al., 1998). It could be argued that this is due to the nature and length of the stimuli (45secs) as the participants may be looking for social/emotional cues whilst listening to the narrative (Alsius et al., 2016). Another study found that the participants focused more on the nose and mouth when sentences were presented in noise (multi-talker babble) again suggesting that the area directly surrounding the mouth is important (Buchan et al., 2008). In the no noise condition when a different talker spoke on every trial, the participants focused on the mouth more compared to when the talker was consistent across trials suggesting talker identity influences where people look (Buchan et al., 2008). Buchan et al. (2008) suggest this is consistent with a strategy in which viewers try to learn the identity of the talker by focusing on the mouth as the physical attributes of the mouth may provide cues about the talker’s voice, which can aid AV integration.

Degrading the visual information and using eye-tracking to see where people look on a face can help to establish which part of the visual stimulus is important for understanding speech in noise. Wilson et al. (2016) found that time spent fixating the mouth of a talker increased as the quality of the visual information increased. Speech perception was still accurate even when high spatial frequency visual information was removed from the face. In contrast, Alsius et al. (2016) degraded both the auditory and visual information using multi-talker babble and visual blurring and found that accurate speech perception was higher when the visual signal was clearer.

Time spent fixating the mouth and eyes also increased as the quality of visual information increased. This shows that individuals may look at the mouth more when there is a benefit of doing so; when the visual stimulus offers better quality speech information.

4.1.3 Aims

Collectively, these studies emphasise the importance of visual information for speech perception. What is unclear is how important fixating a talker's mouth is for AV integration under degraded conditions. The present experiment aimed to clarify how perception of the McGurk effect and eye movements differ in background noise and using degraded visual stimuli. The overall aims were 1) to investigate how perception of the McGurk effect changes when both auditory and visual speech are degraded, 2) to explore eye movements in different levels of white noise and visual blur, and 3) to manipulate fixation cross position as this could have an influence on where people fixate on a face. This could account for some of the inconsistency in the literature in terms of whether fixating the mouth is important.

4.1.4 Hypotheses

It is hypothesised that McGurk responses will increase in auditory noise due to increased influence of the visual modality, but decrease in visual blur. As previous research shows that removing high frequency information is not detrimental to McGurk effect perception, McGurk responses will be reported with some visual blur but will decrease when visual information is severely degraded. Additionally, the McGurk effect will be more likely to be perceived when the participants are fixating the mouth, and this effect may be strongest when a peripheral fixation cross was used as the participants are required to make an eye movement to task relevant areas of the face such as the mouth. Following Gurler et al. (2015), stronger perceivers of the McGurk effect will look at the mouth more than weak perceivers. The results will establish how the weighting of the auditory and visual modalities changes when information from both is suboptimal.

4.2 Method

4.2.1 Design

This experiment used a 3 x 3 x 2 mixed design. The within-subjects factors were Auditory Noise (No noise, Mid noise, High noise) and Visual Blur (No blur, Mid blur, High blur). The between-subjects factor was Fixation Cross position (Central, Peripheral). The dependent variables were McGurk effect perception (proportion of responses which reflect the illusion), defined as responses the participants made that correspond with the non-auditory signal, and dwell time on the mouth (%). A 2 x 6 x 4 design was used for additional analyses with the within subjects factors Congruence (congruent syllables, McGurk syllables), AOI (hair/forehead, left eye, right eye, nose, mouth, forehead/hair) and Talker (Talker 1, 2, 3, 4).

4.2.2 Participants

The participants were 37 students, 5 males and 32 females, aged from 19-48 years old (*M*= 22.35) from Nottingham Trent University. G*power 3.1.9.2 (Faul, Erdfelder, Lang, & Buchner, 2007) was used to determine the sample size needed for a 3-way interaction. An a priori power analysis was conducted, a 3 x 3 x 2 design was specified and a Cohen's *f* of 0.40 (large effect size) was used based on the large effects reported in previous work (Fixmer & Hawkins, 1998), power was 0.95. This analysis determined that a sample size of 30 was needed, more participants were collected than necessary to account for any eye tracking data that may be lost due to poor calibration. The project was approved by the Social Sciences Research Ethics Committee. Students received course research credits for their time. All participants were native English speakers and reported NH, and normal or corrected to normal vision. The participants were selected on the basis of the informal inclusion and exclusion criteria as described in Experiment 1.

4.2.3 Stimuli & apparatus

There were 4 stimuli (1 incongruent syllable + 3 congruent syllables) for each talker and 4 talkers. Talkers were selected based on the results of Experiment 1, stimuli from talkers 1(Token 1), 2 (Token 2), 4 (Token 1) and 6 (Token 2) elicited the illusory percepts most consistently. There were three congruent syllables; /ba:/, /da:/ and /ga:/. Incongruent McGurk pairs were auditory /ba:/ and visual /ga:/ ($A_{BA}V_{GA}$). The 4 stimuli from each talker were presented in 9 different conditions (visual blur: clear, mid blur, high blur x auditory noise: clear, mid, high). There was a total of 144 trials (36 incongruent trials, 108 congruent trials).

The visual blur was created using Gaussian blurring at 40% and 60% in

Premiere Pro v 9.0.0. White noise was created using Matlab and added at two Signal-to-Noise Ratios; -8dB and -20dB. Blur and noise levels were decided upon based on pilot testing (See Appendix B); congruent stimuli (BA, GA, DA) were presented from the 4 talkers in 9 separate levels of auditory noise and visual blur. The participants (*N*=10) were asked to report what syllable they perceived. The noise and blur levels at which correct responses decreased to approximately 50% were chosen to constitute the 'high' level of degradation. This was -20dB for the auditory condition and 60% blur for the visual condition. The data point approximately in the middle of ceiling and poor performance was chosen to represent 'mid' noise. This was -8dB for the auditory condition and 40% blur for the visual condition.

All stimuli were presented at the same sound level (average ~70dB) determined by using a Svantek 977 sound level meter combined with an artificial ear (Brüel & Kjær Type 4153). A 19-inch computer screen was used with a resolution of 1920 x1080 pixels and the stimuli filled 75% of the screen with a visual angle of 37.54°. Stimuli were presented via SMI Experiment Centre and using HD280pro headphones (Sennheiser, Wedemark, Germany). Eye tracking was performed with a RED 500 SMI eye tracker and eye movements were recorded for the duration of each stimulus ~2000ms.

4.2.4 Procedure

Given the results of Experiment 1, and to maintain consistency with other research (Gurler et al., 2015; Paré et al., 2003), a forced choice task was used. The participants sat in front of a desk at ~45cm away from the eye tracker. Before the experiment began, participants were instructed to 'watch and listen closely to the videos' whilst eye movements were recorded. A four-point calibration and validation procedure were performed before each participant began the experiment. Participants were required to watch videos of the talkers and then respond by repeating out loud what they heard from the following choices: BA, GA, DA or THA. Responses were recorded using a Dictaphone and the experimenter later entered the responses into a spreadsheet for analysis. There were 6 practice trials, immediately after each video the 4 choices were displayed on the screen and the participants were prompted to verbally state their choice. During the experimental trials all stimuli were displayed in a randomized order and a fixation cross was displayed. As soon as the participants made an eye movement to the fixation cross, this triggered the stimulus presentation. For half of the participants (*N*= 17) the fixation cross appeared in the centre of the screen and for the other half of the participants (*N*= 16) it appeared in one of four corners of the screen approximately 1 inch away from the corner of the screen (~6° visual angles from the centre of the screen). The corner in which the fixation cross appeared was determined with 25% probability for each corner and randomised between trials.

4.2.5 Analyses

ANOVA was used to analyse the results, where appropriate, the Greenhouse-Geisser correction was applied to correct for violations in assumptions of Sphericity. These instances can be identified by non-integer degrees of freedom. Simple main effects analyses and analyses of interactions were carried out using Bonferroni-corrected t-tests. Unless otherwise stated, for analyses the McGurk effect is defined as any non-auditory response. To analyse the eye-tracking data, six main areas of interest (AOIs) were constructed shown in Figure 4.1 As the face remained static, the AOIs were the same size throughout the video and the mouth AOI was created so it covered the mouth aperture at its widest part. The eye-tracking measure selected for analysis was dwell time which is defined as the sum of all fixations and saccades in a particular AOI for the duration (2000ms) of the stimulus.

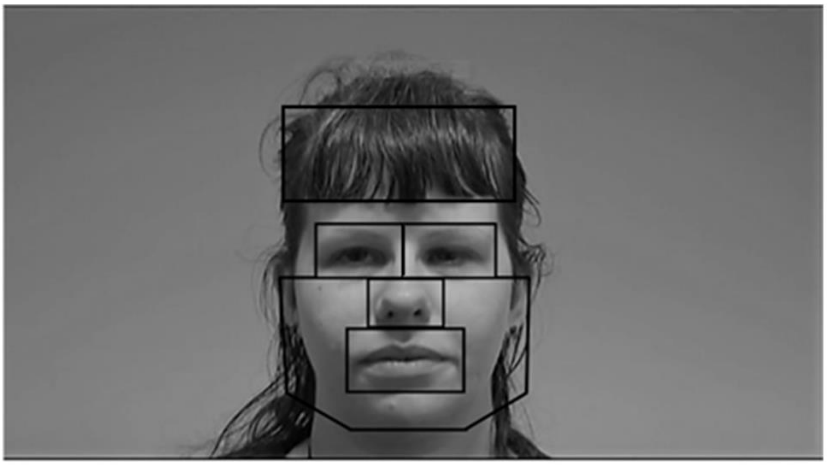


Figure 4.1. Six separate AOIs used encompassing the right eye, left eye, nose, mouth and chin/cheeks and hair/forehead.

4.3. Results

Six participants were excluded after data collection and before analyses were conducted, 4 due to incomplete eye movement data, 1 because of a diagnosis of ADHD and 1 because English was not their 1st language. Therefore, analyses were conducted with 31 participants.

4.3.1 Variability in McGurk effect perception across participants and talkers

Accuracy on congruent trials was 100% in the quiet condition for each congruent stimulus type (BA, GA & DA). Perception of the McGurk effect varied across the participants and talkers, as shown in Figure 4.2. Perception of the McGurk effect ranged from 25-78% (*M*= 60.8%, *SD*= 9.8%) across the participants. Stimuli from different talkers also elicited the McGurk effect by different amounts; for example, the McGurk effect was perceived 86.8% (*SD*= 14.5%) from Talker 2, but just 41.5% (*SD*= 18.1%) of the time from Talker 4.

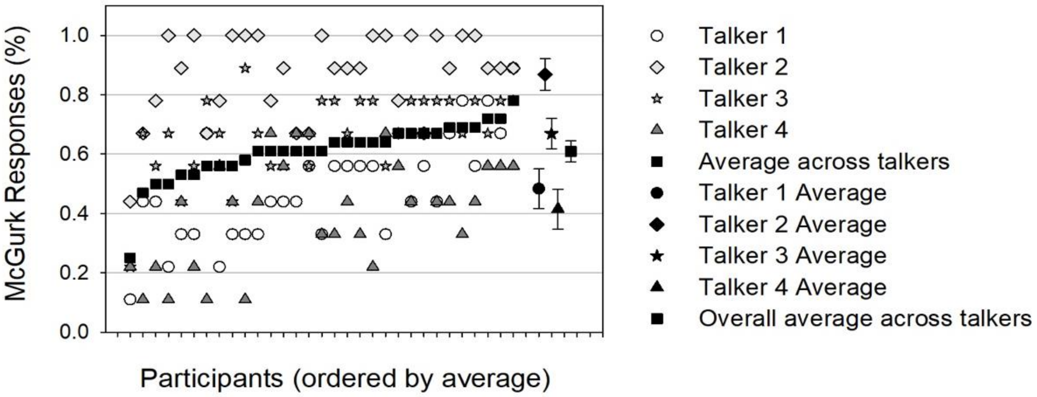


Figure 4.2. Variability in perception of the McGurk effect across the participants and talkers. The participants have been ordered by their average across the 4 talkers. Averages for each talker across the participants are also shown.

4.3.2 Effects of Auditory noise and Visual blur on McGurk responses

The first analysis tested how McGurk responses were affected using a 3 (within subjects factor - Auditory noise: clear, mid noise, high noise) x 3 (within subjects factor - Visual blur: clear, mid blur, high blur) x 2 (between subjects factor- Fixation cross position: central, peripheral) mixed design ANOVA. The analysis showed a significant effect of auditory noise (*F*(2,58) = 87.61, *p* <.001, η^2 = .234) and of visual blur (*F*(1.66,48.07)=104.99, *p* <.001, η^2 = .339), but no effect of fixation cross position (*F*(1,29)=0.02, *p* =.96, η^2 < .000). As shown in Figure 4.3, people made fewer McGurk responses when the auditory signal was clear, and more as the level of auditory noise increased (fewer McGurk responses were made in No noise than in Mid (*p*<.001) or High noise (*p*<.001). There was no significant difference between Mid auditory and High auditory noise (*p*=.0.23). Additionally, more McGurk responses were made when the visual signal was not blurred (more McGurk responses in the No blur than the Mid (p<0.001) or High blur conditions (*p*<.001), and additionally more McGurk responses in Mid blur compared to High blur (*p*=.014). There was a significant interaction between auditory noise and visual blur (*F*(3.21,93.21)=3.66, *p* =.013, η^2 =.017). There was no interaction between auditory noise and fixation type (*F*(2,58,)=1.34, *p* =.27, η^2 = .004), no interaction between visual blur and fixation type (*F*(2,58)=.08, *p* =.92, η^2 < .001), and no interaction between all three, auditory noise, visual blur and fixation type (*F*(4,116)=.67, *p*

=.61, η^2 = .003. The interaction between auditory noise and visual blur seems to have arisen because the number of McGurk responses fell between those of the auditory mid and high noise conditions in the visual clear condition, possibly because McGurk responses almost reached ceiling levels for mid-auditory noise.

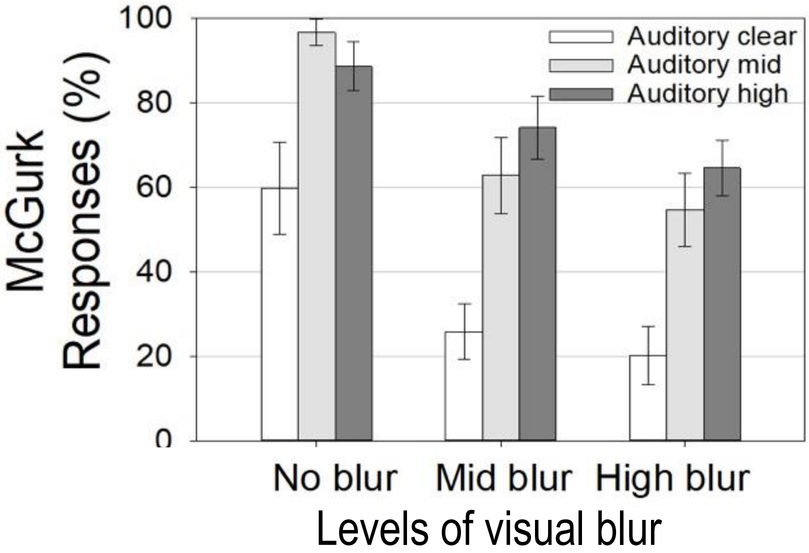


Figure 4.3. McGurk effect perception in auditory noise and visual blur. Error bars denote 95% confidence intervals.

4.3.3 Distribution of eye movements in each Area of Interest (AOI)

Figure 4.4 shows the distribution of eye movements across the different AOIs for each Talker. The pattern of fixations was broadly similar for the different talkers and across Congruent and Incongruent stimuli, with the mouth receiving the most dwell time (overall average 25.9%, *SD* 18.8%), followed by the nose (overall average 17.9%, *SD* 10.1%), followed by the eyes, then the hair/forehead and the chin/cheeks.

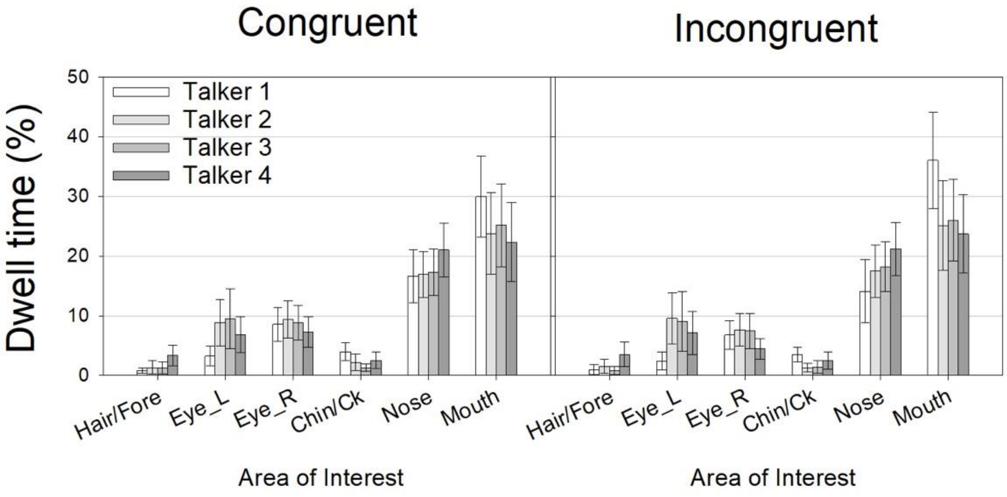


Figure 4.4. Percentage of Dwell time in each Area of Interest according to Congruence (Congruent is the average of three stimuli and Incongruent refers to the single McGurk stimulus) and Stimulus. Error bars represent 95% confidence intervals. The left panel shows data for Congruent stimuli and the right panel shows data for Incongruent (McGurk) stimuli.

A 2 (Congruence) x 6 (AOI) x 4 (Stimulus) ANOVA confirmed that there were significant differences in dwell time according to AOI (*F* (5, 155) = 29.59, *p*<0.001, η^2 = .396). There was additionally a significant interaction between Congruence and AOI (*F* (5, 155) = 10.16, *p*<0.001, η^2 = .002). A comparison of the data in Figure 4.4 (right panel) shows that this was partly driven by dwell times on the mouth being longer for incongruent stimuli (*M* = 27.73%, *SD* = 19.51%) than for congruent stimuli (*M* =25.31, *SD* =18.65%; *t* (31) = 3.71, *p* < .001, *d* = .041). There were additionally significant interactions between AOI and Stimulus (*F* (15, 465) = 10.52, *p* < .001, η^2 = .024) and Congruence, AOI, and Stimulus (*F* (15, 465) = 1.98, *p*= .015, η^2 = .001). As shown in Figure 4.4, the overall pattern of fixations across the different talkers were broadly similar, but there were somewhat different patterns of fixations for the different talkers. For example, Talker 1 (who produced Stimulus 1) elicited more fixations on the mouth than the other stimuli, particularly so when the stimuli were incongruent. The following analyses include just the incongruent (McGurk) stimuli.

4.3.4 Effects of Auditory noise and Visual blur on Dwell times on the Mouth

The next analysis tested dwell time on the mouth using a 3 (within subjects factor - Auditory noise: clear, mid noise, high noise) x 3 (within subjects factor - Visual blur: clear, mid blur, high blur) x 2 (between subjects factor- Fixation cross position: central, peripheral) mixed design ANOVA. There was a significant main effect of visual blur (*F*(1.62,47.03)=11.36, p < .001, η^2 = .042). Figure 4.5 shows that overall, people spent less time fixating the mouth when the visual signal was blurred compared with when it was clear (significantly more time was spent fixating the mouth in the No blur condition than in the High blur (p<.001) or Mid blur conditions (p=.025), but no significant difference between the Mid and High conditions (p=0.10). There was additionally a significant interaction between Visual blur and Auditory noise (*F*(4,116)=3.46, p =.01, η^2 =.007). Figure 5.4 shows that there were different effects of auditory noise depending on how degraded the visual signal was; when the visual signal was blurred, the participants looked at the mouth more as levels of auditory noise increased. There was no main effect of auditory noise (*F*(2,58)=82.78, p =.46, η^2 =.001), no main effect of fixation cross position (*F*(1,29)= .51p =.48, η^2 = .017), no interaction between auditory noise and fixation type (*F*(2,58)=.19, p = .83, η^2 = <.001), no interaction between visual blur and fixation type (*F*(2,58)=1.80, p =.18, η^2 = .007) and no interaction between all three, auditory noise, visual blur and fixation type (*F*(4,116)=.54, p =.71, η^2 = .001).

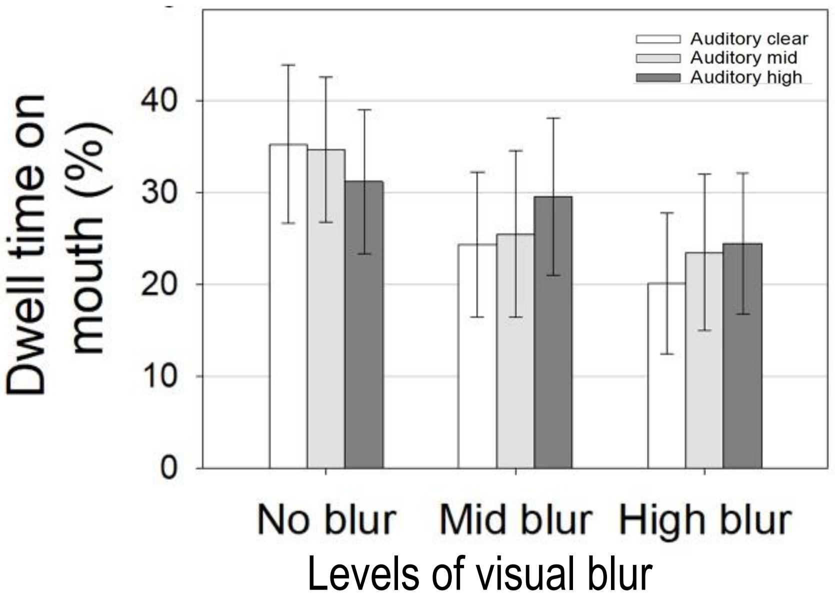


Figure 4.5. Dwell time on mouth (%) in auditory noise and visual blur. Error bars denote 95% confidence intervals.

4.3.5 Dwell time on mouth: Effect of fixation cross position

Figure 4.6 shows the percentage of time the participants spent fixating the mouth according to whether the McGurk effect was perceived, and the location of the fixation cross. A 2 (McGurk effect perception) x 2 (Fixation cross position) mixed ANOVA revealed that significantly longer was spent fixating the mouth when the McGurk effect was perceived (*M* = 29.5%, *SD* = 20.3%) than when it was not (*M* = 25.5%, *SD* = 19.1%; *F*(1,29) = 7.58, *p* = .01, η^2 = .010). There was no significant main effect of Fixation Cross position (*F*(1,29) = .41, *p* = 0.53, η^2 = .014) and no significant interaction between Fixation Cross position and McGurk effect perception (*F*(1,29) = 2.92, *p* = .10, η^2 = .004).

Although there was no significant interaction between Fixation Cross position and McGurk effect perception, Figure 4.6 shows that the main effect of McGurk effect perception seems largely driven by the Peripheral condition. Indeed, there was no significant effect of McGurk effect perception for the Central condition (*t*(14) = - 0.70, *p*= .50, *d* = -.180), but there was for the Peripheral condition (*t*(15) = 3.34, *p*= .004, *d* = -.835).

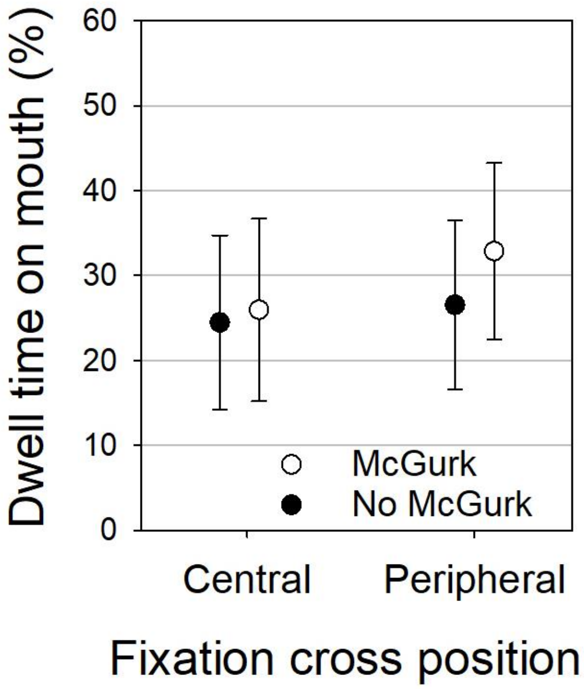


Figure 4.6. Percentage of dwell time on mouth according to McGurk effect perception and Fixation cross position. Error bars denote 95% confidence intervals.

4.3.6 Dwell time on the mouth in strong and weak perceivers

There was large variability in the percentage of time the participants made McGurk responses, ranging from 25 to 78%. Using the traditional classification that strong perceivers experience the McGurk effect on >50% of trials a non-auditory definition of the McGurk effect would dictate that all the participants apart from two were strong perceivers. The average amount the participants perceived the McGurk effect was calculated across stimuli for the non-degraded condition (auditory no-noise and visual no-blur). There was no significant correlation between the average amount the McGurk effect was perceived and the average time spent fixating the mouth *r* = -.169, *p* = .363.

4.4. Discussion

Experiment 2 investigated how perception of the McGurk effect, and accompanying eye movements were affected when speech was presented in auditory noise and visual blur. There was wide variability in perception of the McGurk effect across the participants, ranging from 25-78%. Overall, McGurk responses were made 60.8% of the time. This supports previous findings that the McGurk effect is robust and that vision influences audition in a context when people are presented with incongruent auditory and visual information (Campbell & Massaro, 1997; MacDonald et al., 2000; Thomas & Jordan, 2002). Interestingly, McGurk responses remain at around the 60% level when the auditory and visual signal is subject to the same level of degradation; visual clear + auditory clear = 60%, visual mid blur + auditory mid noise = 63%, visual high blur + auditory high noise = 65%. In terms of the effects of visual blur and auditory noise the hypotheses were confirmed; McGurk effect perception increased in auditory noise and decreased in visual blur. Only when the auditory signal was clear and the visual signal was blurred did McGurk responses fall to under 50%.

According to the PoIE (Meredith & Stein, 1986) it was expected that McGurk responses would increase as auditory noise increases, as unisensory degradation is hypothesized to improve AV integration. However, it was found that when the visual signal was clear McGurk responses peaked in mid auditory noise compared to clear or high noise suggesting that there was an optimum level of auditory noise in which AV integration was advantageous.

As expected the majority of dwell time occurred on the mouth as that is where the speech information is predominantly provided. The second AOI most fixated on was the nose which provides a central location with which to view other features peripherally. This supports studies which found the nose was fixated on more often in noise compared to quiet (Buchan et al., 2007; 2008). The participants looked at the chin/cheek area the least but still sometimes perceived the McGurk effect whilst fixating this area suggesting that they were either processing information from the mouth using peripheral vision or as MacDonald et al. (2000) suggested, that subtle movements of the jaw are sufficient to produce the McGurk effect. Moreover, dynamic articulation of syllables is not just confined to the mouth and includes movements across the whole face (Vatikiotis-Bateson et al., 1998). Whilst this suggests that viewing the mouth is not always necessary to perceive the McGurk effect, the results show that increased McGurk responses are observed when viewers spend more time fixating the mouth. This suggests that fixating the mouth provides richer visual information which contributes to increased illusionary percepts. In support of previous research (Arizpe et al., 2012; Gurler et al. 2015), this effect was driven by those participants who were shown a peripheral fixation cross. Gurler et al. (2015) suggested this could be because the peripheral fixation cross requires the participants to make an eye movement to an area of the face, whereas a central fixation cross encourages the participants to maintain a fixation in the centre of the face and attend to the mouth in their peripheral vision.

Contrary to the findings of Gurler et al. (2015) however, there was no evidence to support the hypothesis that the participants who perceived the McGurk effect more strongly would spend more time fixating the mouth. Again this could be because they were attending to the mouth in their peripheral vision. Pare et al. (2003) found that when the participants' gaze was directed away from the mouth they still reported the McGurk effect suggesting that fixating the mouth is not necessary to perceive the illusion. The present experiment supports this, as it was found that the participants were able to look at the nose, eyes and jaw and still perceive the McGurk effect. As Basu Mallick et al. (2015) point out, categorising participants into strong or weak perceivers of the McGurk effect is determined by the specific stimuli being presented. Therefore, one stimulus may cause a participant to be classed as a strong perceiver and another stimulus may cause them to be classed as a weak perceiver.

Visual blur decreased dwell times on the mouth as expected. The finding of decreased dwell time on the mouth in high levels of visual blur suggests that there was less benefit of the visual information provided by the mouth. Decreased dwell time on the mouth coupled with increased auditory responses in high visual blur suggests that the participants were focusing on the auditory component of the stimulus resulting in reduced McGurk responses.

The findings also demonstrate how AV integration of incongruent information is influenced by degraded conditions. As the McGurk effect, a visually driven illusion, was reduced when the visual signal was degraded and increased when the auditory signal was degraded this supports the modality appropriate hypothesis which states that the senses are weighted based on which modality is the most reliable (Ernst & Bülthoff, 2004; Witten & Knudsen, 2005). However, even when both the auditory and visual information was severely degraded the McGurk effect was still perceived. This suggests that whilst there was a decline in McGurk responses, vision remains influential even when information from both senses is unreliable.

Overall, these findings establish the level of visual and auditory degradation required to inhibit McGurk responses. This is important for understanding how single senses interact when one or both modalities are degraded.

Chapter 5: Experiment 3

5.1 Introduction

Experiment 2 focused on how the McGurk effect and dwell time on the face of a talker are influenced by visual blurring and white noise. Different SNRs of white noise were used as this noise type is appropriate for masking syllables and is akin to background noise experienced in real-world listening environments. Using white noise also allows comparison with previous studies. However, other types of auditory noise should be explored as an additional form of auditory degradation is that experienced by people with hearing impairments. Therefore, Experiment 3 used visual blurring to degrade the visual information and vocoded speech with NH listeners to simulate hearing impairments in noisy environments.

5.1.1 Prevalence of hearing impairments

Given the integral role of speech perception for everyday life, it is concerning that there has been a 12% increase in hearing impairments from 1994 to 2014 (Akeroyd, Foreman & Holman, 2014). The 2014 report on the prevalence of hearing impairments (Akeroyd et al., 2014) suggested that 1 in 12 people aged 18-80 years old suffer with hearing loss across England, Scotland and Wales. Moreover, hearing loss is expected to rise due to the increasing life span of the population (Ciorba, Bianchini, Pelucchi & Pastore, 2012). This growth in hearing impairments is detrimental due to the importance of hearing for communication as well as leisure, for example, listening to music. Hearing impairments can also have a negative impact on an individual's social life causing issues such as low self-esteem, loneliness and depression (Ciobra et al., 2012). Therefore, understanding how individuals utilise visual information to improve speech perception is important, especially for the future, in order to improve people's quality of life.

5.1.2 Hearing impairments and Cochlear-implants

Hearing impairment can be present from birth (congenital) or develop with age. Hearing impairments can be categorised as conductive, sensorineural, or mixed hearing loss, which is a combination of conductive and sensorineural. Conductive hearing loss occurs as a result of damage to the outer or middle ear caused by infection, a perforated eardrum or an abnormal bone structure in the middle ear. This type of hearing loss can be treated with surgery or hearing aids depending on the underlying cause.

The most prolific cause of hearing impairment occurs when the inner hair cells are damaged, resulting in sensori-neural hearing loss. This type of hearing impairment is often permanent as hair cells are unable to proliferate. Hair cells can be damaged for many reasons including; unsafe levels of noise, illness (e.g. meningitis) and as a result of taking certain medications (Woulters, McDermott & Francart, 2015). It’s estimated that 613,000 adults across England and Wales have severe to profound deafness (Raine, 2013). Cochlear-implants can be used to treat severe-profound sensori-neural hearing loss through replicating the function of the ear and partially restoring hearing. Woulters et al., (2015) report that 80,000 children have received cochlear-implants worldwide. Including both children and adults, over 300,000 cochlear-implants have been fitted worldwide (National Institutes of deafness and other communication disorders, 2017).

It is important to understand how a cochlear implant works as this affects the auditory information experienced by the user. Implants are comprised of a microphone, speech processor and transmitting coil which are situated behind the pinna (Loizou, 1998). The internal processor is implanted behind the ear and the electrodes are inserted directly into the cochlea. The exact placement of the electrodes varies across individuals and is determined by which parts of the anatomical structures remain intact (Dorman, Loizou, Fitzke & Tu, 1998). The speech processor transforms acoustic vibrations into electrical impulses which are sent to the electrodes which in turn stimulate the auditory nerve (Rubinstein, 2004). The part of the cochlea that is stimulated depends on the frequency of the signal, as in normal hearing (NH), high frequencies produce activity at the base and low frequencies produce activity near the apex of the cochlea. To achieve this, speech input is divided into different frequency bands, also termed channels. The amount of channels varies according to the specific implant used, however, the fewer channels available the more spectral resolution is reduced. Spectral information refers to frequency based features of the voice which can be used to identify pitch. Fewer channels mean that cochlear-implant users may have access to less spectral information compared to listeners with NH. Fast fluctuations in the speech signal are also omitted as Figure 5.1 shows, these fast fluctuations in amplitude over time are referred to as temporal fine structure (TFS) cues and are important for pitch perception (Moon & Hong, 2014). Therefore, whilst the cochlear implant is able to partially restore hearing the speech information provided does not match the hearing of listeners with NH.

Third party copyright material removed

Figure 5.1. Cochlear implant encoding with four channels. This shows how information is lost through encoding. Panel 1 depicts the original speech signal ‘Sa’, 2 shows the signal divided into frequency bands from (top to bottom) high to low frequency, 3 shows the envelopes extracted – broad amplitude fluctuations over time and TFS removed, 4 shows the pulses generated which correspond to the envelope (Loizou, 1998).

Despite the limitations in cochlear implant processing, some adults with cochlear-implants can achieve a very high level of performance, especially in quiet listening situations. Gantz, Woodworth, Abbas, Knutson and Tyler (1993) reported that cochlear implant users were able to identify up to 96% of words in sentences. Dorman and Loizou (1997, 1998) found that some cochlear implant users with a six-channel implant matched the performance of listeners with NH who were presented speech via six-channels. This suggests that for some, cochlear-implants can offer a high level of speech intelligibility.

However, there is large variability in the success of cochlear-implants depending on the recipient, this is due to several factors including the amount of time individuals were deaf before receiving an implant and their speech perception abilities prior to implantation (UK cochlear implant group, 2004). The performance of the user may also vary depending on the amount of time which has elapsed since the implant was fitted (UK cochlear implant group, 2004). Over 200 cochlear implant users were tested on their ability to identify key words in sentences over a 9-month period (Gantz et al., 1993). It was found that performance improved over time for all users, however performance across individuals ranged from 20% to >80%.

Gantz et al. (1993) also found that after 9-months performance ranged from 0% to 96% when identifying words in sentences and 0% to 46% for words only. This suggests that whilst there can be a vast improvement for some users, others still have relatively poor performance with a cochlear implant.

5.1.3 Vocoded speech

Due to the substantial variability in performance across cochlear-implant users, vocoders have been utilised in research with listeners with NH to simulate the filtering process of cochlear-implants (Dorman, Loizou & Rainey, 1997; Rosen Faulkner & Wilkinson, 1999). Vocoders filter speech through channels in a similar way to cochlear implant processing meaning that spectral and temporal information is diminished. Figure 5.2 depicts the process for creating noise-vocoded speech. First the speech signal is filtered into separate frequency ranges. Second, the amplitude envelope is extracted and smoothed, these envelopes are then used to modulate a carrier signal, and finally information in the channels is recombined (Davis, Johnsrude, Hervais-Adelman, Taylor & McGettigan, 2005).

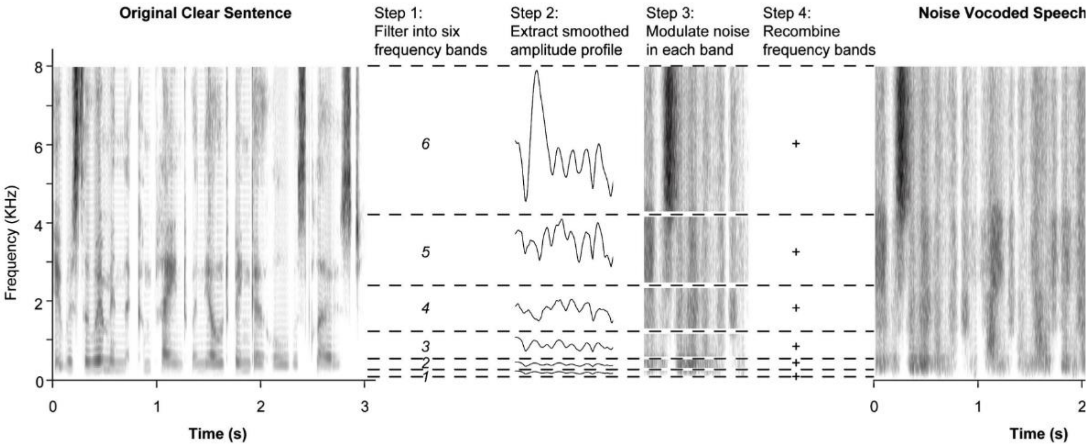

Original Clear Sentence
Step 1: Filter into six frequency bands
Step 2: Extract smoothed amplitude profile
Step 3: Modulate noise in each band
Step 4: Recombine frequency bands
Noise Vocoded Speech
Frequency (KHz)
8
6
4
2
0
6
5
4
3
2
1
+
Time (s)
0
1
2
3
Time (s)
0
1
2

Figure 5.2. Noise-vocoded speech with six channels. From Lexical information drives perceptual learning of distorted speech: Evidence from the comprehension of noise-vocoded sentences, by M. H. Davis, I.S. Johnsrude, A. Hervais-Adelman, K. Taylor and C. McGettigan, 2005, *Journal of Experimental Psychology: General*, *134*(2), p.222. Copyright 2005 by American Psychological Association. Reprinted with permission.

Conducting research using listeners with NH and using vocoders is helpful in many ways as there is often age related cognitive decline associated with hearing loss including deficits in memory and language (Lin et al., 2013). Lin et al. (2013) found that older adults with hearing loss experienced faster cognitive decline compared to older adults with NH which could make it difficult to determine whether performance on a task is due to poor speech intelligibility or cognitive decline. In addition, there is substantial variability amongst cochlear-implant users (Gantz et al., 1993). Collectively, this makes it difficult to determine the factors which contribute to reduced speech intelligibility in cochlear-implant users. Therefore, studies using vocoders and listeners with NH can be informative as there is reduced variability and cognitive decline meaning the effects of noise on speech intelligibility can be more easily isolated.

There are a number of parameters that can be varied when vocoding speech. Vocoders can have different numbers of channels (Davis et al., 2005), and the envelope cut-off frequencies can be varied (Souza & Rosen, 2009). Most vocoder studies tend to use 8-channels of information, this is because although modern cochlear-implants have around 16-24 channels (e.g. Nucles-22), users of cochlear-implants are seldom able to use more than 8 channels of information (Fishman, Shannon & Slattery, 1997). This could be due to a number of reasons including the number of electrodes used (Fishman et al., 1997), and auditory nerve survival (Skinner et al., 2002). Another parameter of vocoders is the cut-off frequency for the temporal fine-structure. This is typically set to 160Hz, Apoux and Bacon (2008) found that performance improved as cut-off frequency increased but that performance was the same for 400Hz compared 160Hz. Finally, either sine-wave or noise carriers can be used to modulate the envelopes.

5.1.4 The challenges of listening in noise for people with normal hearing and people with hearing impairments

Everyday conversation is often hindered by background noise meaning that less information is available in the acoustic signal, resulting in more errors when identifying speech (Peelle, 2018). This can be exacerbated for people with hearing impairments for whom the auditory signal is already degraded. Background noise refers to general noise in the environment or noise produced by multiple talkers which masks speech information (Rubinstein, 2004).

Masking effects make the perception of speech in noise difficult. Brungart, Simpson, Ericson and Scott (2001) outline two types of masking. Informational masking is where both the auditory signals from the target talker and the competing talker are perceptible but the listener is incapable of separating the two streams; the listener can hear two sources of information, but they may get confused between the streams. This is in contrast to energetic masking which is where the target speech signal and competing signal share similar temporal and frequency cues rendering the target signal imperceptible. Brungart et al. (2001) found that listeners with NH found it more difficult to decipher speech when the target talker was masked by noise compared to competing talkers. Listeners were better at segregating speech from multiple talkers if the target voice was a different sex to that of the competing signal. Furthermore, if the competing signals were from multiple same sex talkers, performance was better compared to a competing signal from one different sex talker. These findings show how differences in competing talkers relative to the target talker can influence successful segregation of speech.

Understanding speech in background noise can be challenging for all listeners, but it poses a particular challenge for users of cochlear-implants. According to Wouters, McDermott and Francart (2015) cochlear implant users need an SNR which is 15 dB higher than listeners with NH in order to reach 50% correct performance. One reason for this is the loss of TFS which can contribute to poor pitch perception (Qin & Oxenham, 2003). In order to decipher speech from competing noise, the individual must be able to segregate sounds from competing sources, a process known as streaming (Oxenham, 2008). This can be difficult with poor pitch perception as different sounds may be perceived as one (Moon & Hong, 2014).

Studies have also shown that speech perception in noise is difficult when listeners with NH are presented with vocoded speech. Qin and Oxenham (2003) manipulated the number of channels available in vocoded speech when speech was also masked by multiple talkers. The ability of listeners with NH to identify speech was reduced, even with eight channels available. Stickney, Nie and Zeng (2005) compared cochlear-implant users with listeners with NH and manipulated the number of channels in noisy and quiet conditions. In the quiet condition, cochlear-implant users were able to identify sentences correctly 70% of the time which matched the performance of listeners with NH when speech was presented with ten channels. In contrast, when speech was masked with a competing talker, cochlear-implant performance declined to 10% matching the performance of listeners with NH with four channels. This demonstrates how lack of TFS can make it difficult to segregate speech from competing talkers. Overall, cochlear-implant users' speech perception can decline considerably when speech is masked by noise, and the performance of listeners with NH can be reduced to match that of those with hearing impairments by reducing the number of channels available.

As cochlear-implant users struggle to identify speech in noise, it is important to understand how visual speech cues aid cochlear-implant users with communication. Previous research has successfully used vocoded speech with listeners with NH in order to inform the design of training programmes which aim to improve speech perception in cochlear-implant users (Rosen, Faulkner & Wilkinson, 1999; Stacey & Summerfield, 2007; Stacey et al., 2010). Pilling and Thomas (2011) found that the participants were able to better understand speech which was degraded using an 8-channel noise-vocoder when they were trained using AV speech compared to when they were trained with auditory speech alone. Recently, Blackburn, Kitterick, Jones, Sumner and Stacey (2019) investigated how the use of visual speech information varied across different listening situations and talkers when listening to speech in noise. The participants were listeners with NH who were presented with clear speech or speech processed by an 8-channel sine-wave vocoder. They found that people received more benefit from visual speech information when speech was vocoded rather than clear (consistent with Stacey, Kitterick, Morris & Sumner, 2016), and the amount of visual speech benefit varied according to the intelligibility of the target talkers and the number of talkers in the background noise. This research is informative as it shows that the amount that people benefit from visual speech information can vary according to listening situations and task demands.

5.1.5 Different types of vocoder: Sine-wave vs Noise-vocoded

Two types of carrier are most frequently used to create vocoded speech; sine-wave (tone) or noise carriers. The perceptual experience of the listener can differ depending on the type of carrier used. Noise-vocoded speech has the effect of quieting the voice or adding noise whereas sine-vocoding results in flattened pitch (Dorman, Natale, Butts, Zeitler & Carlson, 2017).

Several studies have compared different types of vocoder with listeners with NH in an attempt to establish which vocoder provides the best simulation of cochlear-implants. Dorman et al. (1997) compared the ability to identify sentences, consonants and vowels, which were presented via either a sine-vocoder or a noise-vocoder. They found that there were no differences in performance depending on vocoder type (Dorman et al., 1997). Similarly, Xu (2016) compared sine-wave speech and natural speech (sentences) which was either sine-vocoded (tone) or noise-vocoded. Sentence recognition was similar for both sine- and noise-vocoded speech however, performance was slightly higher for both noise-vocoded speech types. Whitmall, Poissant, Freyman and Helfer (2007) found that sine-vocoded syllables and sentences were identified more than noise-vocoded speech in both quiet and noise. The authors suggest that a limitation of the noise-vocoder is that the temporal fluctuations of the noise carriers hamper the temporal fluctuations needed for identifying speech. Laneau, Moonen and Wouters (2006) explored the suitability of the noise-vocoder for modelling cochlear-implant users' pitch perception in several experiments. They compared cochlear-implant users, and listeners with NH using two types of noise-vocoders. They found that the ability of listeners with NH to discern important speech cues including place pitch cues and temporal cues was poorer compared to cochlear-implant users. This suggests that whilst noise-vocoders are useful for modelling cochlear-implants, listeners with NH may not match the performance of cochlear implant users and therefore the results should be interpreted with caution.

Compared to noise-vocoding, sine-vocoding is considered more perceptually similar to cochlear-implants, due to the tonal quality of the output (Dorman et al., 1997), indeed cochlear-implant users reported that single channel stimulation sounds like beeps rather than noise (Gonzalez & Oliver, 2005), Overall, this suggests that the sine-vocoder may be preferable to the noise-vocoder.

5.1.6 Limitations of vocoder simulations

Despite the usefulness of vocoder studies for understanding speech perception in users of cochlear-implants, recent research suggests that vocoded speech does not replicate the subjective experience of sound in cochlear-implant users (Dorman et al., 2017). One study asked cochlear-implant users with unilateral deafness, to listen to different types of vocoded speech and rate them on a scale of zero to ten in terms of the similarity to the speech delivered from their cochlear implant. Sentences were presented with one of the following: noise-vocoding, sine-vocoding, frequency shifted sine-vocoding or band-pass filtered natural speech. The number of channels was also manipulated for noise- and sine-vocoded sentences in increments of 2, ranging from 4-12 channels. Individuals reported that the vocoded stimuli did not match that of their cochlear-implants. However, some individuals found that increasing the band-pass filtering of natural speech produced a voice that sounded like the equivalent of speech heard through their cochlear implant, whilst others reported that decreasing the band-pass filtering to produce a muffled voice, was more akin to the speech delivered through their implants. This finding suggests that the subjective experience of listening through a cochlear implant varies across individuals, therefore, one type of vocoder may only simulate speech degradation for some cochlear implant users and not others.

A final caveat is that studies testing listeners with NH do not reflect the substantial variability observed in studies with cochlear implant users. This is due to individual differences in people who use cochlear-implants and differences in the design of the cochlear-implants (see Section 5.1). However, the advantage of vocoders is that studies can be carried out with listeners with NH, to assess factors that are difficult to isolate in cochlear-implants, for example, manipulating the number of channels. Results from vocoding studies can inform further research with cochlear implant users, with a view to improving cochlear implant functionality and the performance of people with cochlear-implants.

The implications of these studies are that the type of vocoder used should be taken into account when making comparisons with listeners with NH and cochlear implant users. Research (Dorman et al., 1997; Laneau et al., 2006) suggests an overall advantage of sine-vocoded speech suggesting that sine-vocoders are the most appropriate for simulating cochlear-implants.

As discussed in Section 5.1.2 people with profound deafness can have their hearing partially restored by cochlear-implants (CIs) however; CIs do not restore NH but deliver a signal that is temporally and spectrally degraded meaning they often struggle to understand speech in noise. Research with CI users suggests they benefit from visual information and may be more adept at AV integration compared to people with NH (Rouger et al., 2007). In conjunction with this, CI users perceive the McGurk effect more often compared to NH listeners (Stropahl, Schellhardt, Debener, 2016). This benefit of visual information and increased perception of the McGurk effect could be due to CI users' tendency to look at the mouth more compared to people with NH (Mastrantuono, Saldaña & Rodríguez-Ortiz, 2017). People with CIs might look at the mouth more in order to help them get more information from the visual signal, when the auditory signal is degraded. This can be tested in normal-hearing listeners by using vocoded speech (Shannon et al., 1995) which simulates the speech processing involved in a CI (Dorman et al., 1997; Rosen et al., 1999). Using NH listeners is advantageous as there are several barriers to understanding speech intelligibility in CI users. Firstly, research shows there is large variability in CI performance (UK cochlear implant group, 2004) and, secondly, CI users often also have age related cognitive deficits linked to hearing loss. Therefore, vocoded speech provides a method of simulating hearing loss whilst avoiding these confounds. This could also shed light on how visual cues aid hearing impaired listeners with understanding speech in noise. Results of the experiments could contribute towards designing training for CI users to assist them with speech perception.

Vocoding degrades the speech in two ways through extensive blurring of the frequency information presented, and rapid fluctuations in amplitude over time are removed. This impairs the understanding of speech in quiet and in noisy environments (Qin & Oxenham, 2003). Studies find that there is more benefit from seeing the face of a talker when speech is vocoded compared to clear speech (Blackburn et al., 2019; Stacey et al., 2016). However, these studies did not use eye-tracking so it is unclear what specifically about the visual information was important.

Using vocoding to simulate hearing impairments and presenting speech in white noise and visual blur will help to understand how visual information is used in difficult listening situations.

Experiment 2 showed that fixating the mouth in peripheral vision or fixating the jaw was sufficient for the McGurk effect to occur however in more difficult listening situations such as this, fixating the mouth may be even more informative. As CI users adopt a strategy where they look at the mouth (Mastrantuono et al., 2017), using vocoded speech with NH listeners may result in more time spent fixating the mouth compared to Experiment 2. In addition, white noise masks syllables whereas vocoding omits temporal and spectral information from the speech signal meaning that reduced access to important speech cues may result in increased reliance on the visual information. Gaussian blurring is also perceptually more similar to vocoding than white noise as this type of visual degradation minimises speech cues through the removal high frequency visual information. Both types of degradation have been found to reduce speech intelligibility (Munhall et al., 2004; Shannon et al., 1995).

5.1.7 Aims

Experiment 3 used vocoded speech to simulate the information provided by a cochlear implant. The overall aims were 1) to investigate how AV integration changes when speech is subject to both auditory and visual degradation 2) to explore eye movements in different levels of white noise with vocoded speech and visual blur, and 3) to manipulate fixation cross position as this could have an influence on where people fixate on a face as experiment 2 found that fixation cross position influenced fixating on the mouth. Vocoding degrades the speech signal both spectrally (by blurring across frequency) and temporally (by removing rapid fluctuations in amplitude over time). Vocoding also allows for more ease of comparison with the visually degraded (Gaussian blurred) stimuli. As CI users often struggle to understand speech in noise it is important to study vocoded speech to understand how eye movement strategies can aid AV integration. This would elucidate which parts of the face are important in different noise contexts. In addition, it is useful to note that hearing impaired listeners have other age related cognitive deficits, and it is helpful to conduct initial experiments with NH listeners to inform future research with hearing impaired listeners.

5.1.8 Hypotheses

Previous research shows that vocoding impairs speech perception (Qin & Oxenham, 2003). It is expected that people will look at the mouth more in challenging listening conditions when speech is vocoded as well as presented in white noise compared to when the only source of noise is from vocoded speech. It is hypothesised that the results of Experiment 2 will be replicated and perception of the McGurk effect will increase as auditory noise increases and decrease as visual blur increases.

5.2 Method

The same equipment and procedure were used as in Experiment 2. The same participants as Experiment 2 completed Experiment 3; the participants completed Experiments 2 and 3 in a counterbalanced order. Six participants were excluded from both experiments due to incomplete eye movement data meaning analyses were conducted with 31 participants.

The stimuli were presented with the addition that the auditory signal was vocoded as well as presented in white noise (visual blur: clear, mid blur, high blur x auditory, vocoded, vocoded with mid white noise, vocoded with high white noise). The stimuli were vocoded prior to the experiment in Matlab (Mathworks) using an 8-channel vocoder. The stimuli were band-pass filtered into 8 adjacent frequency bands spaced equally on an equivalent rectangular bandwidth frequency scale between 100 Hz and 8 kHz (Glasberg & Moore, 1990) using Finite Impulse Response filters. The temporal envelope of each filter output was extracted using the Hilbert transform and used to modulate a sine wave at the central frequency value of the filter. The eight sine waves were then summed. Pilot testing, as described in Experiment 2, revealed that for vocoded speech performance fell to approximately 50% correct at an SNR of -9dB (high noise condition). An SNR of 0dB fell between this and ceiling performance levels for vocoded speech, so was chosen for the Mid auditory noise condition. Visual blurring was at 40% (mid) and 60% (high).

5.3 Results

5.3.1 Variability in McGurk effect perception across participants and stimuli

Accuracy on congruent trials was 100% in the quiet condition for each congruent stimulus type (BA, GA & DA). McGurk effect perception varied across the participants, ranging from 55 to 92% (*M*= 72.9%, *SD* = 9.7%). There was also large variability in the perception of the McGurk effect across the talkers, as Figure

5.3 shows. With Talker 2 the McGurk effect was perceived 92.3% of the time (*SD* 25.8%), while with Talker 1 the McGurk effect was perceived 60.5% of the time (*SD* 48.9%).

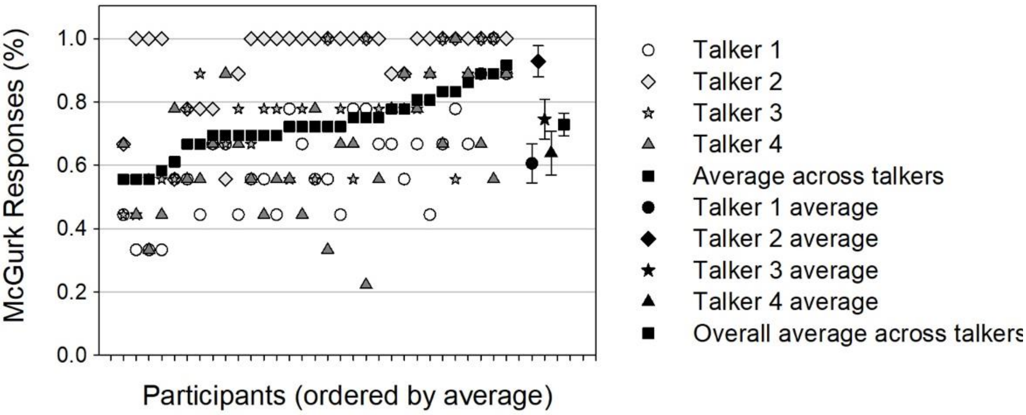


Figure 5.3. Variability in perception of the McGurk effect across the participants and talkers. The participants have been ordered by their average across the 4 talkers. Averages for each talker across the participants are also shown.

5.3.2 Effects of Auditory noise and Visual blur on McGurk Responses

An analysis was conducted which tested McGurk effect perception using a 3 (within subjects factor - Auditory noise: vocoded speech in clear, mid noise, high noise) x 3 (within subjects factor - Visual blur: clear, mid blur, high blur) x 2 (between subjects factor- Fixation cross position: central, peripheral) mixed design ANOVA. There was a significant main effect of Visual blur (*F*(1.53,44.43)=41.46, *p* < .001, η^2 = .264), indicating that fewer McGurk responses were made when the visual stimulus was blurred than when the stimulus was clear (Mid visual blur vs No blur: *p*<.001, High visual blur compared to No blur: *p*<.001, Mid blur vs High blur: *p*=.393). Figure 5.4 shows that when the visual signal was clear there was a high level of McGurk responses, regardless of auditory degradation. There was no effect of auditory noise (*F*(2,58)=2.23, *p* = .11, η^2 = .008), no effect of fixation cross position (*F*(1,29)= .91, *p* = .34, η^2 = .030), no interaction between auditory noise and visual blur (*F*(4,116)=.97, *p* = .43, η^2 =.008), no interaction between auditory noise and fixation type (*F*(2,58)=1.54, *p* = .31, η^2 =.004), no interaction between visual blur and fixation type (*F*(2,58)= 1.62, *p* = .20, η^2 =.010), and no interaction between all three auditory noise, visual blur, and fixation type (*F*(4,116)=1.59, *p* = .18, η^2 =.023).

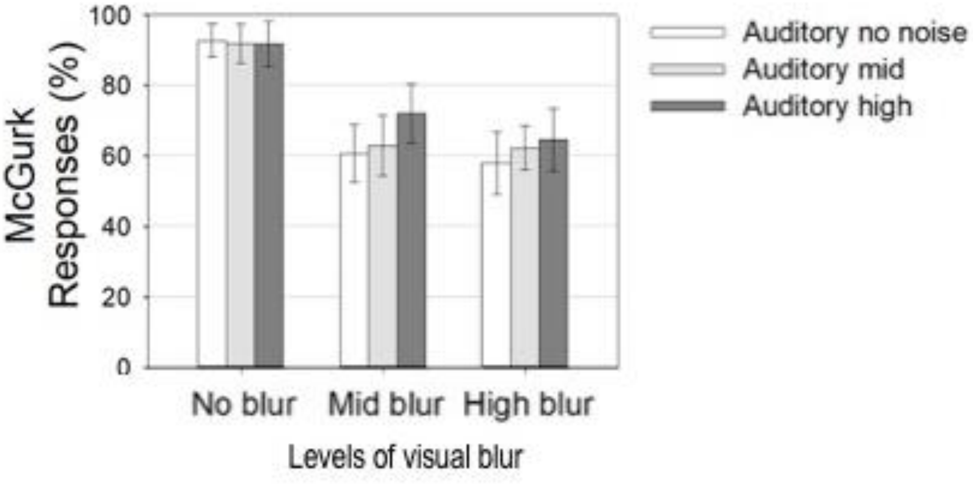


Figure 5.4. McGurk effect perception in auditory noise and visual blur. Error bars denote 95% confidence intervals.

5.3.3 Distribution of Dwell time in each Area of Interest (AOI)

Figure 5.5 shows the dwell time (as defined in Experiment 2) within each AOI for each stimulus. As with Experiment 2, the mouth received the most dwell time, followed by the nose and then the eyes. The differences in dwell time across AOIs was significant, as expected (*F* (5, 155) = 27.73, *p*<.001, η^2 =.397). There were small variations in this pattern according to which talker the participants were viewing and whether the stimuli were congruent or incongruent, but this pattern was broadly consistent across stimuli. There was nevertheless a significant interaction between Congruence and AOI (*F* (5,155) = 3.33, *p*<0.01, η^2 = .001); slightly more time was spent fixating the mouth and less time was spent fixating the eyes when stimuli were incongruent than when stimuli were congruent (Figure 5.3). Additionally, a significant interaction between AOI and Stimuli (*F* (15, 465) = 5.46, *p*< .001, η^2 = .009) was found because the amount of dwell time in each AOI varied slightly for the different stimuli. For example, more time was spent on the mouth of Talker 1 than the mouth of other stimuli.

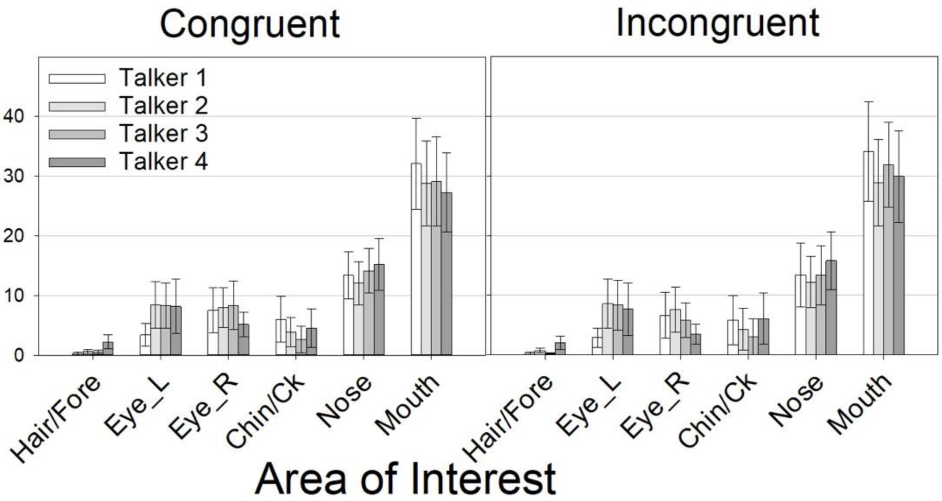


Figure 5.5. Percentage of Dwell time in each Area of Interest according to Congruence (Congruent is the average of three stimuli and Incongruent refers to the single McGurk stimulus) and Stimulus. Error bars represent 95% confidence intervals. The left panel shows data for Congruent stimuli and the right panel shows data for Incongruent (McGurk) stimuli.

5.3.4 Effects of Auditory noise and Visual blur on Dwell times on the mouth

Dwell time on the mouth was tested using a 3 (within subjects factor - Auditory noise: vocoded speech in clear, mid noise, high noise) x 3 (within subjects factor - Visual blur: clear, mid blur, high blur) x 2 (between subjects factor- Fixation cross position: central, peripheral) mixed design ANOVA. There was a significant main effect of Visual blur (*F*(1.6,8.15)=5.22, *p* = .009, η^2= .015; post-hoc comparisons showed that the participants looked at the mouth less (*M* =29.06%, *SE* =3.85) when there was a high level of visual blur compared with when there was no visual blur (*M*= 35.45%, *SE*= 3.64, *p*=.032), shown in Figure 5.6. There was no effect of auditory noise (*F*(2,58)=2.90, *p* = .063, η^2 = .004), no effect of fixation cross position (*F*(1,29)=.04, *p* = .83, η^2 = .002), no interaction between auditory noise and visual blur (*F*(4,116)=.94 *p* = .14, η^2 = .003), no interaction between auditory noise and fixation type (*F*(2,58)=.77, *p* = .46, η^2 = .001), no interaction between visual blur and fixation type (*F*(2,58)= .39, *p* = .67, η^2 = .001), and no interaction between all three auditory noise, visual blur, and fixation type (*F*(4,116)=.33, *p* = .85, η^2 = .001).

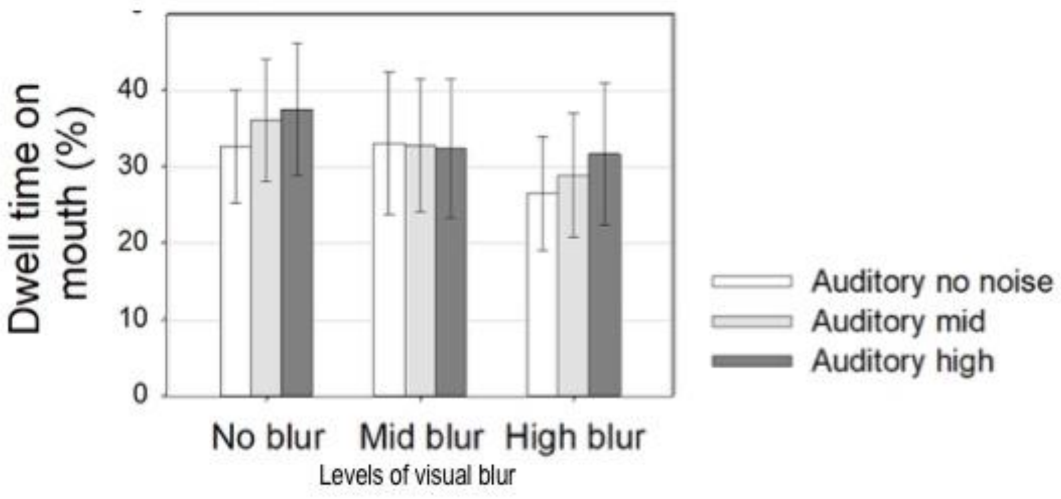


Figure 5.6. Dwell time on mouth (%) in auditory noise and visual blur. Error bars denote 95% confidence intervals.

5.3.5 Dwell time on mouth: Effect of fixation cross position

A mixed ANOVA was conducted with 2 (within subjects factor – McGurk perception: perceived or not perceived, x 2 (between subjects factor – fixation cross: central, peripheral). Figure 5.7 shows that the participants spent longer fixating the mouth when they perceived the McGurk effect (*M* = 33.0%, *SD* = 20.4%) compared to when they did not (*M*= 30.7%, *SD* = 22.1%), but this was not significant *F*(1,29) = 3.03, *p* = .092, η^2 = .012). There was no significant main effect of Fixation cross position (*F*(1,29) = 0.65, *p* = .62, η^2 = .022 and no significant interaction between Fixation Cross position x McGurk effect perception (*F*(1,29) = 0.77, *p* = .39, η^2 < .000).

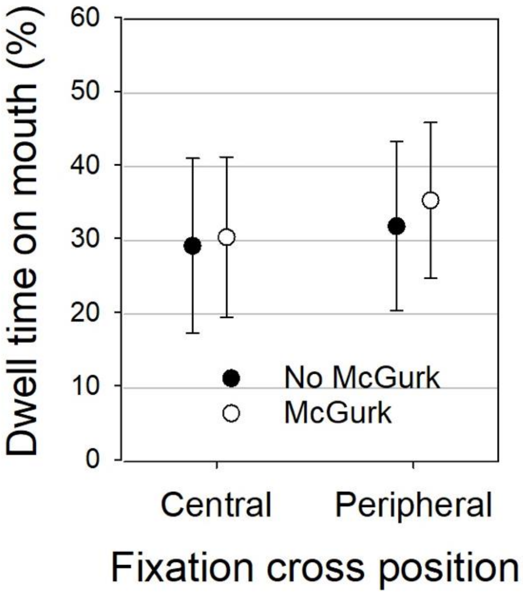


Figure 5.7. Percentage of dwell time on mouth according to McGurk effect perception and Fixation cross position. Error bars denote 95% confidence intervals.

5.3.6 Dwell time on mouth in strong vs. weak perceivers

The relationship between strength of McGurk effect perception and time spent fixating the mouth in quiet was examined and found no significant correlation *r* = -.047, *p* = .81.

5.4 Discussion

Experiment 3 aimed to establish how looking at the mouth of a talker influences the McGurk effect to gain insights into AV integration when the stimuli are degraded by visual blur, vocoding and white noise. Consistent with the results from Experiments 2, variability in the McGurk effect was demonstrated with the effect being perceived between 55-92% across the participants. On average, across all noise levels, the McGurk effect was perceived 72.6% of the time, which is higher than the 60.8% reported in Experiment 2. The higher visual influence found in this experiment is likely due to the poorer intelligibility of the auditory signal when speech is vocoded. Only in one condition does perception of the McGurk effect fall to below 50%; for the Auditory clear x High visual blur condition with a central fixation cross, where McGurk perception falls to 48.3%.

Consistent with the results of experiment 2, as visual blur increased, McGurk effect perception decreased as well as dwell time on the mouth. This provides support for previous research which suggests that people only look at the mouth when there is additional benefit from the visual information (MacDonald et al., 2000; Wilson et al., 2016). Dwell time in each AOI was similar to Experiment 2 as the participants spent the majority of time focused on the mouth, followed by the nose. When incongruent and congruent stimuli were compared the participants spent more time fixating the mouth when the stimuli were congruent.

Overall, the participants spent 31.0% of the time fixating the mouth region, which is slightly higher than, but comparable to, the 27.7% in Experiment 2. Unlike the findings of Experiment 2 people did not spend longer fixating the mouth when the McGurk effect was perceived compared to when it was not perceived. As in Experiment 2 there was no relationship between time spent fixating the mouth and the McGurk effect when strong and weak perceivers were compared.

Unlike Experiment 2, there were no effects of auditory noise on McGurk effect perception or dwell time on the mouth. One explanation is that the vocoded speech was difficult for listeners with NH to understand, therefore, the inclusion of white noise applied to the vocoded stimuli may have had no additional effect. Using different talkers in the stimulus set also makes it more challenging to identify vocoded speech (Loizou, Dorman & Tu, 1999). Dwell time in each AOI was explored in relation to talker, it was found that more time was spent fixating the mouth for Talker 1 compared to the other talkers.

5.4.1 Comparison of findings from experiments 2 & 3

To date it has not been well understood how auditory and visual information interact under degraded conditions, or how beneficial fixating a talker’s mouth is for AV integration in these conditions. The present experiments investigated how the relative signal strengths of modalities in multisensory task settings affect the extent of multisensory integration as well as dwell time on the face of a talker. AV integration was measured by perception of the McGurk effect in different levels of auditory noise and visual blur. This is relevant for people with both auditory and visual impairments and for understanding how AV integration is influenced when information from one or more modalities is degraded.

Overall, across Experiments 2 and 3, it was found that AV integration was robust; the McGurk effect, which was defined as a change in the auditory percept, averaged 60.8% in Experiment 2 and 72.6% in Experiment 3. Only when visual information was degraded and the auditory signal was presented with no noise did the frequency of the McGurk effect fall to below 50%. According to the Principle of Inverse Effectiveness (Meredith & Stein, 1986) we would expect McGurk responses to increase as auditory noise increases, as unisensory degradation is hypothesized to improve AV integration. The results support this hypothesis; when there was noise in the auditory signal perception of the McGurk effect increased and people also looked more at the mouth. In Experiment 2 it was found that when the visual signal was not blurred McGurk responses peaked in mid auditory noise compared to no noise or high noise. As expected, adding blur to the visual signal decreased perception of the McGurk effect and also dwell times on the mouth.

Fixation cross position was manipulated to clarify if the starting position influences where people look on a face. Overall, whilst fixation cross position did not influence dwell time on the mouth, in the peripheral fixation cross condition the participants were more likely to look at the mouth more when the McGurk effect was perceived. This suggests that eye-movement measures may only reveal effects when purposeful eye movements need to be made to areas of interest, as otherwise the participants may rely on information they can obtain in their peripheral vision. The finding in Experiment 2 that visual blur had a greater effect in the peripheral fixation cross condition than in the central fixation cross condition supports this conclusion.

Contrary to previous research (Gurler et al., 2015) stronger perceivers of the McGurk effect did not look more at the mouth. One explanation is that strong perceivers were able to make use of the visual information from other areas of the face. Indeed, the finding that the McGurk effect was still evident when faces and voices were severely degraded suggests that viewers were still able to glean enough visual information to produce the effect. In high visual blur when the mouth was barely discernible, the McGurk effect was still perceived (in Experiment 2 20% of the time for no auditory noise, and 58% of the time for mid auditory noise). Although viewers looked at the mouth less, focusing on other areas of the face was sufficient for the McGurk effect to be perceived. The findings provide support for previous work measuring eye movements in visual blur (Alsius et al., 2016; Wilson et al., 2016) suggesting that viewers look at the mouth more when there was a benefit of doing so; when high spatial frequency information was intact.

As the second most fixated AOI was the nose, the participants could have also viewed the mouth peripherally. Moreover, dynamic articulation of syllables is not just confined to the mouth and includes movements across the whole face (Vatikiotis-Bateson et al., 1998). Whilst this suggests that fixating the mouth is not always *necessary* to perceive the McGurk effect, the results show that increased McGurk responses are observed when viewers spend more time fixating the mouth. This suggests that fixating the mouth provides richer visual information which contributes to increased illusory percepts. The finding that higher levels of auditory noise led to more time fixating the mouth supports the suggestion that in challenging listening situations people look more at the most useful aspect of the face to obtain visual speech information. This is also supported by the finding that more time was spent fixating the mouth when the stimuli were incongruent than when they were congruent.

The findings presented here serve to resolve some of the contradictions regarding whether or not fixating the mouth is important for McGurk perception. When the visual signal is not blurred and the mouth is fixated this increases the likelihood of the McGurk effect being perceived. Accordingly, one would expect people to receive greater benefit from visual speech information when the visual signal is not degraded and the mouth is fixated. While the McGurk effect is still perceived to some extent when the visual signal is blurred, the results suggest that if the visual signal is blurred then people will receive less benefit from visual speech information, and they will disengage from looking at the mouth. The ability to integrate auditory and visual information varies across individuals and populations including older adults (Sekiyama et al., 2014) and people with hearing impairments (Tye-Murray, Spehar, Sommers et al., 2016). Therefore, future research should continue to examine AV integration with both auditory and visual degradation with these populations as they may rely more on visual signals.

The findings also demonstrate how AV integration of incongruent information is influenced by degraded stimulus presentations. The McGurk effect, a visually driven illusion, was reduced when the visual signal was degraded and increased when the auditory signal was degraded. This supports the modality appropriate hypothesis which states that the senses are weighted based on which modality is the most reliable (Ernst & Bülthoff, 2004; Witten & Knudsen, 2005). However, even when both the auditory and visual information were severely degraded the McGurk effect was still perceived. This suggests that whilst there was a decline in McGurk responses, vision remains influential even when information from both senses is unreliable.

5.4.2 Conclusion

The McGurk effect is a widely reported illusion that occurs when auditory and visual information is conflicting, and is still perceived even when the visual signal is severely degraded. Fixating the mouth is not strictly necessary for the McGurk effect to occur but the McGurk effect increases when the visual signal is clear and the mouth is fixated. This suggests the possibility that the best strategy for greater AV integration in auditory noise may be to fixate the mouth. Future work should examine this possibility outside of the context of perception of the McGurk effect, such as when listeners are presented with conversational speech in background noise.

Chapter 6: Experiment 4

6.1 Introduction

Experiments 2 and 3 used the McGurk effect to investigate how AV integration is influenced by fixating the mouth and what part of the visual stimulus is important. The findings showed that both dwell time on the mouth, and perception of the McGurk effect increase when the visual signal is clear. The present experiment aimed to build on this work by using different speech stimuli (words) presented in noise. These words were presented in Auditory Only (AO) and AV conditions, and visual benefit was calculated by the difference in performance between these conditions. The primary goal of this experiment was to establish whether similar results were found as in Experiments 2 and 3 when a different measure of AV integration was used as this establishes whether both measures share the same mechanism for AV integration. Examining eye movements in this context would help to elucidate which part of the visual stimulus is important and how eye movements influence visual benefit.

6.1.1 Aims & hypotheses

6.1.1.1 Aim a: To explore visual benefit for accuracy and visual benefit for RT in response to word stimuli when these are degraded through auditory noise and visual blur.

As discussed in Section 1.9.2, previous research shows that NH listeners benefit from seeing a talker's face when listening in background noise (Sumby & Pollack, 1954). The behavioural advantage is reflected in higher accuracy scores for AV speech than auditory only speech (AV-AO) and will be referred to as visual benefit. A highly related measure is visual gain which several studies have attempted to quantify (Altieri & Wenger, 2013; Altieri & Townsend, 2011; Sumby & Pollack, 1954). Altieri and Townsend (2011) used a paradigm in which the participants were required to make speeded responses to one of 8 different words (Mouse, Job, Tile, Gain, Shop, Boat, Page, and Date) under AO and AV conditions. This procedure allowed them to calculate (1) visual gain for accuracy (VG_A) using the formula AV-AO/1-AO and (2) visual gain for reaction time (RT, VG_RT) calculated as AO-AV. VG_A measures the gain in accuracy scores from seeing the face of a talker relative to the auditory information alone. VG_RT represents the influence of visual information on processing speed.

Altieri and Townsend (2011) measured visual gain in quiet and at two SNRs; -12dB and -18dB. On AV trials when speech was presented in quiet, AV integration was more accurate and RT was faster compared to the noise conditions. VG_A and VG_RT increased as auditory noise increased (-18 dB) meaning there was an advantage of being able to see a talker's face when speech was presented in higher levels of noise. Visual gain also decreased as the auditory signal improved. Using a similar paradigm, Altieri and Wenger (2013) found that when the auditory signal was clear there was little or no visual gain observed for accuracy or RT. Visual gain increased as noise increased and RT was faster in the AV condition compared to the AO condition. Overall these results suggest a clear auditory signal is sufficient for speech perception and that visual information provides an advantage when the auditory signal is degraded in noise.

The effects of visual degradation in AV integration are described in Section 1.10.4 It is likely that visual degradation will also negatively affect visual benefit, consistent with the results of Experiments 2 and 3.

Hypothesis a) It was expected that visual benefit for accuracy and RT will increase when words are presented in auditory noise but will decrease when words are degraded using visual blur.

6.1.1.2 Aim b: To replicate the results of Experiments 2 and 3 which showed that dwell time on the mouth decreased in visual blur, and to see if auditory noise influences fixations on the mouth.

Monitoring gaze whilst the participants view stimuli in noise can help to elucidate which part of the visual stimulus is important for visual benefit. This literature was reviewed in Section 1.10.1. Blackburn (2019) also found that time spent looking at the mouth was a significant predictor of visual benefit. The present experiment investigated whether similar results were found with degraded word stimuli as were found with the degraded incongruent stimuli used in Experiments 2 and 3.

Hypothesis b) It was expected that the mouth will be fixated less in visual blurring as the quality of visual information decreases. Based on the results of Alsius et al (2016) people will fixate the mouth more in the presence of auditory noise to compensate for the reduced reliability of the auditory signal.

6.1.1.3 Aim c. To investigate how eye movements differ according to individual differences in visual benefit measured by AV-AO.

The next aim was to establish whether people who received more visual benefit also looked more at the mouth. Alsius et al. (2016) examined participants' visual gain in different levels of auditory noise and visual blur using words and sentences. They found that accuracy improved more for some individuals compared to others when the visual stimulus was clear, suggesting that these individuals were more adept at extracting visual information. Based on these results, the participants were divided into high visual gain (HVG) and low visual gain (LVG) groups. The HVG group also spent more time looking at the mouth of the talker compared to the LVG group for the word stimuli but not for sentences.

Hypothesis c) In line with the findings from Alsius et al. (2016), it was expected that more visual benefit will be found in people who fixate the mouth more.

6.1.1.4 Aim d: To investigate whether there is a relationship between visual benefit for words in noise and McGurk perception

The final main aim was to explore whether there was a relationship between two different measures of AV integration; visual benefit for words and McGurk perception. The final main aim was to explore whether there was a relationship between two different measures of AV integration; visual benefit for words and McGurk perception. Van Engen et al. (2017) highlight the assumption held by many researchers who use the McGurk effect that visual gain and McGurk effect perception are related, in that they both share the same mechanism for AV integration. As the McGurk effect is a visually driven illusion, this would suggest that people who perceive the McGurk effect more frequently (strong perceivers) may also be better at extracting visual information and therefore experience more visual gain than people who perceive the McGurk effect less frequently (weak perceivers). Van Engen et al. (2017) wanted to test the hypothesis that strong perceivers of the

McGurk effect would also experience more visual gain; they found that when sentences and incongruent stimuli were presented in noise, visual gain for sentences was not predicted by perception of the McGurk effect. This suggests that further research is needed to establish the relationship between the McGurk effect and visual gain to better understand the influence of visual speech information in different contexts, and if both measures reflect similar AV integration processes. One explanation for Van Engen et al's (2017) results is that speech processing may vary according to the type of speech stimuli used. Van Engen et al. (2017) compared sentences with incongruent syllables, sentences are more complex and offer "richer contextual cues" (Alsius et al., 2016) compared to words or syllables. Therefore, there may exist a relationship between the ability to identify words in noise and McGurk perception as both words and syllables are limited in contextual cues. Measuring both McGurk perception as well as visual benefit for words is advantageous as it provides a further measure of AV integration using congruent speech, which is more akin to natural speech in everyday conversation compared to incongruent stimuli. Furthermore, using word stimuli has the advantage over sentence stimuli as words are shorter and therefore appropriate for measuring reaction time.

Hypothesis d) It was expected that visual benefit will increase as McGurk effect perception increases.

6.2 Method

6.2.1 Design

The study used a within-participants factorial design. The independent variables were auditory noise with three levels (clear, mid, high) and visual blur with three levels (clear, mid, high). Three dependent variables were analysed separately; visual benefit for accuracy, RT gain, and the percentage of time looking at the mouth. Visual benefit for accuracy was calculated using AV-AO and for RT was calculated using AO-AV. The participants completed an initial learning block (64 trials) at the start of the experiment with auditory only stimuli to learn the key placement and to prevent them from looking down at the keyboard to preserve the eye-tracking data. This block was repeated (64 trials) after two of the test bocks had been administered, this was to remind the participants of the key placement. For the main experiment words were presented in AV, VO and AO blocks which were counterbalanced across the participants. AV words were presented in 3 levels of auditory noise (clear, mid, high) and 3 levels of visual blur (clear, mid, high). There were 144 AV trials (16 words (8 x 2 talkers) x 9 noise conditions). The AO block (48 trials) consisted of words presented in clear, mid and high auditory noise. The visual only block (48 trials) also included silent videos of the talkers uttering the words in three levels of visual blur. Finally, there was a McGurk block (80 trials) which was always included at the end of the experimental session, this block included congruent syllables and incongruent McGurk syllables al presented in clear listening conditions. All measures and conditions are outlined in figure 6.1.

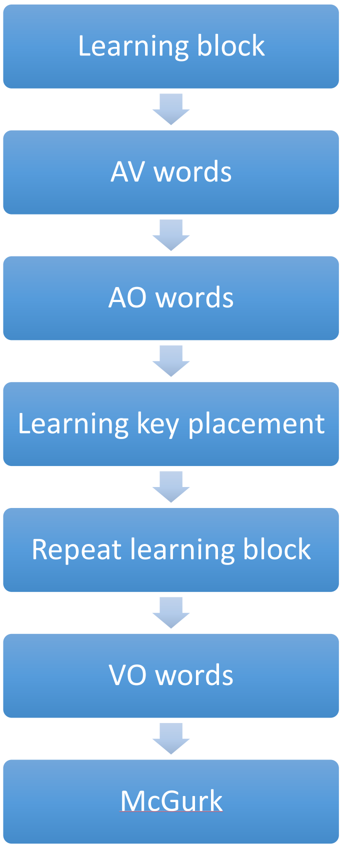


Figure 6.1. Measures and conditions used. Each participant received the blocks in consecutive order, except for the AV, AO and VO blocks which were counterbalanced across the participants.

6.2.2 Participants

G*power 3.1.9.2 was used to determine the sample size (Faul et al., 2007) needed for a 2-way interaction. A 3 x 3 within-subjects design was specified and a Cohen's *f* of 0.40 was used which represents a large effect size in line with effect sizes reported in previous studies with a similar design (e.g. Alsius et al., 2016). Power was specified as standard (0.8). This analysis determined that a minimum sample size of 20 was needed. More participants were recruited than necessary as it was assumed some participants' data would have to be excluded due to poor calibration with the eye-tracker. Therefore, a total of fifty-one participants were recruited which matched the sample size from Alsius et al (2016). The participants were selected on the basis of the informal inclusion and exclusion criteria as described in Experiment 1 and were recruited from Nottingham Trent University. Nine were excluded (4 due to poor eye-tracker calibration, 2 could not learn the words required to complete the task, 2 pressed the wrong keys and 1 reported a diagnosis of dyslexia). A total of 42 participants were included in the final data set aged 18 to 31 (*M* = 20.14 years, *SD* = 2.63 years). The participants were informed of their rights and informed consent was obtained in line with Nottingham Trent University's ethical procedures.

6.2.3 Stimuli & apparatus

The same eye tracking apparatus and set-up were used in Experiment 4 as in Experiments 2 and 3.

6.2.3.1 Words stimuli.

The stimuli consisted of videos and sound files of two female talkers articulating the following monosyllabic words: Mouse, Job, Tile, Gain, Shop, Boat, Page, and Date. These words were chosen from previous work (Altieri & Townsend, 2011; Altieri & Wenger, 2013) which used stimuli from the Hoosier Audio-visual Multi-talker database (Sheffert, Lachs & Hernandez, 1996), as the talkers in these stimuli have American accents stimuli were recreated for the present experiment using English talkers. The short length of the words and the small set size were appropriate for collecting RT data where fewer response options are advantageous (Hick, 1952). All the word stimuli were edited so that the video files were exactly three seconds in duration, there were 144 stimuli in the AV block (2 talkers x 8 word stimuli = 16, x 9 levels of noise: 3 Auditory x 3 visual). The same noise levels were applied as used in experiments 2 and 3, -8 and -20 SNRs for the auditory stimuli and 40% and 60% Gaussian blur for the visual stimuli. Forty-eight stimuli were used in the visual only block (2 talkers x 8 word stimuli = 16, x 3 levels of visual noise) and audio only block (2 talkers x 8 word stimuli = 16, x 3 levels of auditory noise). All word stimuli were presented through SMI Experiment Centre software via the eye-tracker, responses were recorded automatically. Experiment builder software (E-prime v2.0) was used to present the incongruent stimuli in clear listening conditions. There was also a learning block which was used so that the participants could learn the key placement, this consisted of the 16 clear auditory only stimuli which were used in the main experiment, these were repeated four times (64 trials in total). The learning block was repeated, in total each stimulus 8 times (8 x 16 stimuli = 128 trials in total).

6.2.3.2 Incongruent stimuli

The incongruent stimuli were ~2 seconds in duration. There were 3 types of stimuli, audio-visual (video with sound), visual only (video without sound) and audio only (sound file). There were 8 stimuli in the McGurk perception block (2 talkers x 4 stimuli BA, GA, DA, $A_{BA}V_{GA}$) which were presented 10 times (in line with Basu-Mallik et al., 2015) making 80 trials. The same talkers were used for the word stimuli as the incongruent stimuli.

6.2.4. Procedure

As the experiment involved eye tracking, the procedure started with a learning block in which the participants were required to learn which words corresponded with which numbers on the keyboard, this was to prevent the participants looking down at the keyboard during the task and to preserve the eye movement data. An 8 option forced choice task was used (in line with Altieri & Wenger, 2013), the words Mouse, Job, Tile, Gain, Shop, Boat, Page, and Date corresponded to numbers 1-8 on the keyboard. For the main task three blocks were presented, these consisted of audio-visual (AV), visual only (VO) and audio only (AO). A subsequent learning block was always presented after the first two blocks to aid the participants in remembering how the responses were mapped onto the keyboard. McGurk perception was included as an additional measure so that visual benefit scores from words in noise could be compared to frequency of the McGurk effect. The final block consisted of incongruent stimuli intermixed with the congruent syllables BA, GA, and DA. This block was always presented last so that any illusory responses perceived did not interfere with responses to the words. The AV (144 trials, 8 stimuli x 9 noise repeated), VO (8 stimuli x 3 noise repeated = 48 trials) and AO (48 trials) blocks were counterbalanced across the participants. Accuracy and RT were measured on all blocks. Eye-tracking was only used on the AV and VO blocks. Before each trial a peripheral fixation cross was presented in one of the four corners of the screen, the video or sound would then play and the participants responded by pressing one of the 8 numbered keys as fast as possible. The experimenter then manually triggered the next trial. The whole experiment lasted 60 minutes.

6.2.5 Data analysis

Figure 6.2 shows the distribution of scores for the learning block for the data available from 39 participants. As an 8 option forced choice task was used chance was calculated at 13%. Whilst some participants' overall accuracy (*N*=2) fell between 6-25% on test trials these participants scored 100% in the AV clear condition suggesting that they had accurately learnt the key placement therefore they were included in the main analyses.

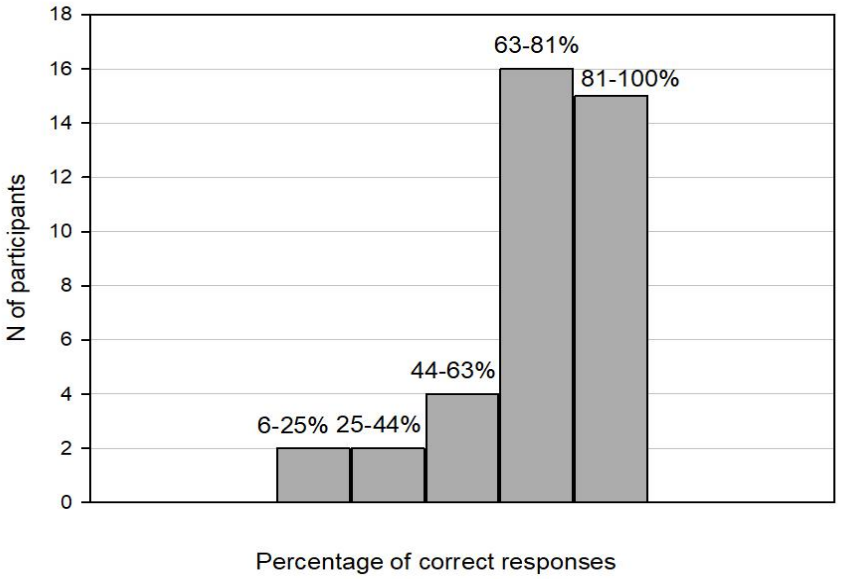


Figure 6.2. Distribution of responses to the learning block (*N*=39)

Eye movement data was divided into three areas of interest (AOIs) which spanned the eyes, nose and mouth consistent with Alsius et al. (2016). The AOIs depicted in Figure 6.3 differed from experiments 2 and 3 which also included separate AOIs for each eye, an AOI for the hair/forehead, and an AOI for the chin/cheeks. The latter two AOIs were omitted as dwell time was negligible in these areas and the eye AOIs were combined for easier analysis. The percentage of dwell time in each area was calculated, this includes every eye movement from the first fixation in a particular AOI to the last. The following calculation was applied to the accuracy data before the main analyses were carried out to access visual benefit (AV-A), this reflects how much accuracy scores improve with the addition of seeing a talker's face compared to the auditory only condition. The formula for visual gain (AV-A/1-AO) could not be applied to the data as accuracy data in the AO condition for some participants was at ceiling, and as Altieri and Wenger (2013) noted when scores are at ceiling visual gain scores become redundant. The RT data were screened for outliers and RT that was +3SD away from the mean were excluded. RT data only included responses to correct trials only. RT data is often positively skewed meaning it violates the assumptions of parametric tests. Several alternatives were considered and ruled out: a) non-parametric tests have low statistical power b) transforming the data is not ideal as it often does not resolve skewness and does not prevent Type 1 errors (Ratcliff, 1993); c) Miller (1998) notes that if the median is used with unequal trials across conditions (as in the present study) this can artificially inflate differences between conditions. Therefore, mean RT was used with ANOVA for consistency with previous work and because ANOVA is generally robust to non-normal data. The following calculation was applied to the RT data (AO-AV). As not all participants had data for all parts of the experiment (due to technical problems with the experimental software), at the start of each analysis the number of participants included in the analysis is indicated.

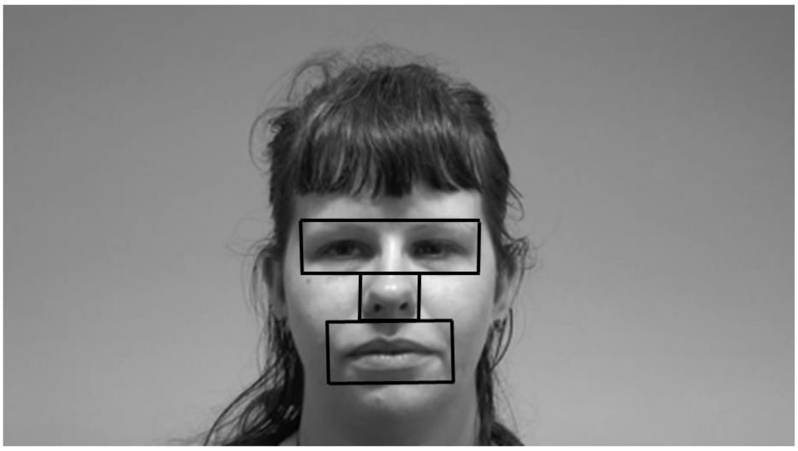


Figure 6.3. Three AOIs used encompassing the eyes, nose and mouth.

6.3 Results

6.3.1 Summary of accuracy and RT

The data shown in Table 6.1 are the overall mean percentage of words correctly identified and the mean RT averaged across all words and conditions in each modality.

Table 6.1

Mean accuracy and RT across modality collapsed across noise types

Accuracy			
Modality	*Mean*	*CI lower*	*upper*
AO	71%	67	74
VO	35%	31	38
AV	74%	71	78
Reaction time			
AO	1665	1573	1757
VO	1753	1690	1823
AV	1905	1823	1986

Table 6.1 shows that individuals were faster and more accurate overall in the AO condition and slower and less accurate in the VO condition. A one-way repeated measures ANOVA with DV accuracy and IV modality with 3 levels AO, VO, AV

was significant *F*(2,82) = 226.29 *p* <.001, η^2 = 0.736. Pairwise comparisons showed that there was no significant difference in accuracy between AV trials and AO trials (*p* = .194), accuracy was lower on VO than AO (*p* <.001) or AV trials (*p* <.001). One-way repeated measures ANOVA with DV RT and IV modality with 3 levels AO, VO, AV was significant *F*(1.63,66.95) = 8.10 *p* < .05, η^2 = 0.083. Pairwise comparisons showed that RT was faster on AO than VO (*p* <.001), or AV trials (*p* =.008). There was no significant difference between RT on VO trials and AV trials (*p=1.00*).

6.3.2 Aim a: How is visual benefit affected by auditory noise and visual blur?

6.3.2.1 Visual benefit for accuracy (N = 42)

Visual benefit was calculated using AV-AO on accuracy scores. Figure 6.4 shows that there was increased visual benefit when words were presented in high auditory noise compared to clear or mid auditory noise indicating that there was a benefit to seeing the talker's face when the auditory stimulus was harder to understand.

A 3 x 3 repeated measures ANOVA was conducted with DV visual benefit and IVs auditory noise (3 levels: clear, mid, high) and visual blur (3 levels: clear, mid, high). There was a significant effect of visual blur *F*(2,82) = 10.07, *p*<.001, η^2 = 0.011, auditory noise *F*(2,82) = 87.24, *p*<.001, η^2 = 0.298 and a significant interaction *F*(4,164) = 13.62 , *p*<.001, η^2 = 0.027.

Positive numbers indicate that in high auditory noise visual benefit increased suggesting that the addition of seeing the face aided in deciphering the auditory signal, data are plotted according to the auditory conditions to illustrate this in Figure 6.4. To examine the interaction three separate one-way ANOVAs were conducted on visual benefit scores in each level of auditory noise with Bonferroni correction. There was no significant difference in visual benefit in the auditory clear condition across the three levels of visual blur *F*(2,82) = .239 , *p*= .788, η^2 = 0.001 and no significant difference in visual benefit in the auditory mid noise condition across three levels of visual blur *F*(2, 82) = .299 , *p*= .742, η^2 = 0.001. Visual benefit differed significantly in the auditory high condition and increased as the clarity of the visual stimulus increased *F*(2,82)=25.396, *p*<.001, η^2 = 0.165 Pairwise comparisons showed that there was significantly more visual benefit in the clear visual condition compared to the mid blur condition (*p*=.004) and the high blur condition (*p*<.001). There was also significantly more visual benefit in the mid blur condition compared to the high blur condition (*p*=.003). This suggests that visual information was of most benefit when high frequency visual information was included on the face.

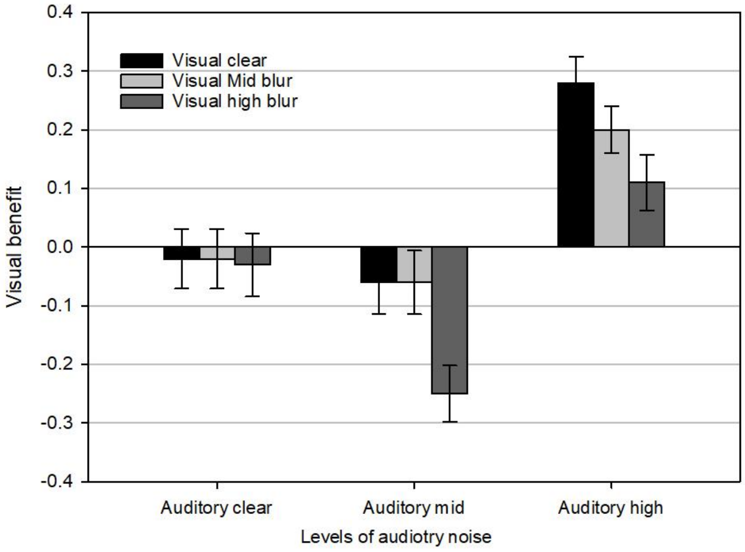


Figure 6.4 Visual benefit (accuracy) for words in auditory noise and visual blur, error bars represent 95% confidence intervals.

6.3.2.2 Visual benefit for RT (N = 42)

RT gain was calculated by using AO-AV, positive numbers indicate an AV advantage. Figure 6.5 indicates that RT was faster for AV words than AO words when the visual stimulus was clear and presented in clear auditory and high auditory noise. In general, the inclusion of visual information slowed responses. Large standard deviations indicate that there was substantial variability across the participants. The participants received more RT gain and were faster when the visual stimulus was clear, but the amount of RT gain received was minimal (<40ms).A 3 x 3 repeated measures ANOVA was conducted on RT and with auditory noise (3 levels: clear, mid noise, high noise) and visual blur (3 levels: clear, mid blur, high blur). There was a significant effect of visual blur *F*(2,82) = 6.57, *p*=.002, η^2 = 0.006 and no significant effect of auditory noise *F*(1.6, 67.02) = 0.56, *p*=.917, η^2 < 0.000

and no significant interaction *F*(2.8, 116.42) = .717, *p*=.536, η^2 =0.001. Pairwise comparisons for visual blur showed that there was significantly more RT gain in the visual clear condition compared to the mid (*p*=.005) or high blur (*p*=.014) condition. There was no significant difference between the mid and high blur conditions (*p*=1.00).

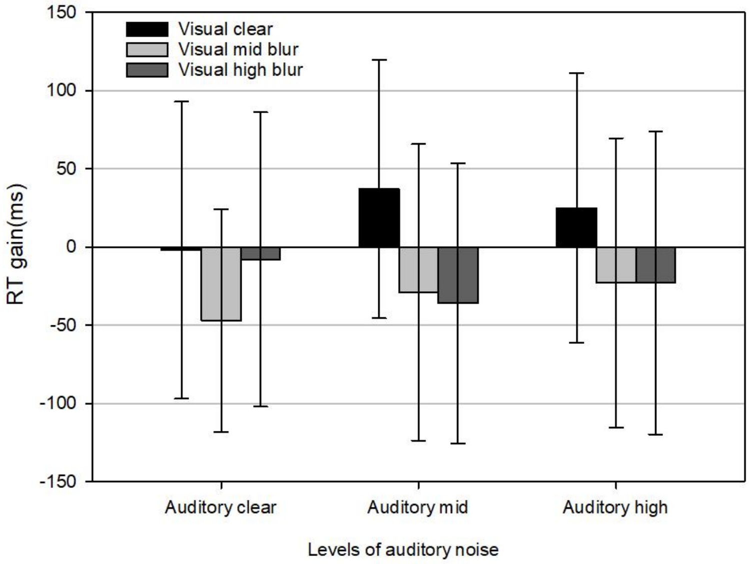


Figure 6.5 RT gain in auditory noise and visual blur, error bars represent 95 % confidence intervals.

6.3.3 Summary of eye movement data (*N* = 42)

On VO trials the participants looked at the mouth the most compared to the eyes and nose whereas for the AV trials the participants looked more at the eyes compared to the mouth and nose, illustrated in Figure 6.6. One-way ANOVA with dwell time in each AOI (eyes, nose and mouth) was conducted separately for each modality for the clear condition only and showed a significant effect of AOI for the VO trials *F*(1.39,56.88) = 11.98, *p* = <.001, η^2 = 0.181. Pairwise comparisons showed that the participants looked significantly more at the mouth than nose (*p*<.001) and eyes (*p* =.004). There was a significant effect of AOI for AV trials *F*(2, 82) = 9.033, *p*= <.001, η^2 = 0.124 as the participants looked at the eyes more than the mouth (*p*=.025) and nose (*p*=.001).

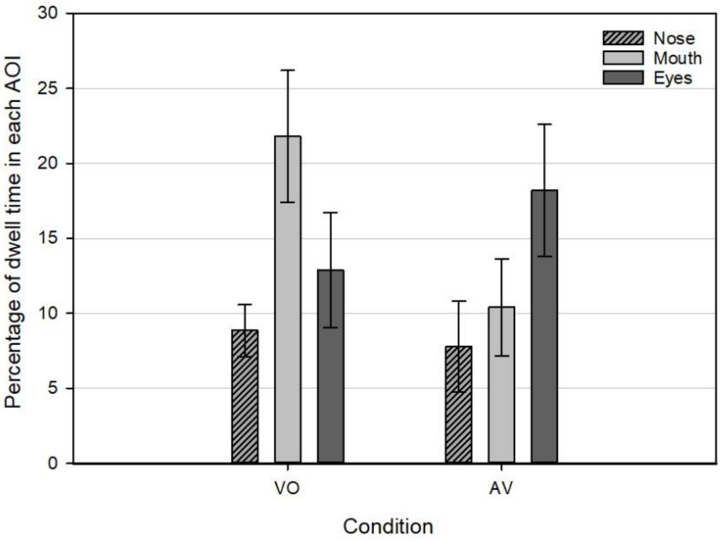


Figure 6.6 Average dwell time on the mouth in each AOI according to clear VO and clear AV conditions, error bars represent standard 95% confidence intervals

As the largest percentage of dwell time was spent on the mouth in the VO conditions a one-way ANOVA was carried out on the mouth across 3 levels of visual blur (clear, mid blur, high blur), there were no differences in dwell time on the mouth in different levels of visual blur $F(2,82) = 1.29$, $p=.279$, $\eta^2 = 0.003$. Overall, in the absence of auditory information the participants were not influenced by visual blur. Mouth dwell time on AV trials according to the different levels of blur is presented in Section 6.3.4.

6.3.4 Aim b: How is mouth dwell time influenced by noise? (*N* = 42)

To facilitate comparison with experiments 2 and 3, eye movements in the different levels of noise were explored.

Figure 6.7 shows the participants looked at the mouth less in high visual blur compared to the mid blur and clear conditions and more in high auditory noise compared to mid auditory noise. A 3 x 3 repeated measures ANOVA was performed with DV mouth dwell time and IVs auditory noise (3 levels: clear, mid noise, high noise) and visual blur (3 levels: clear, mid blur, high blur). There was a significant effect of visual blur $F(2,82) =5.85$, $p= .004$, $\eta^2 = 0.012$, and auditory noise $F(2,82) = 5.55$, $p=.005$, $\eta^2 = 0.003$ and no significant interaction $F(4,164) = 1.82$, $p= .127$, $\eta^2 = 0.001$.

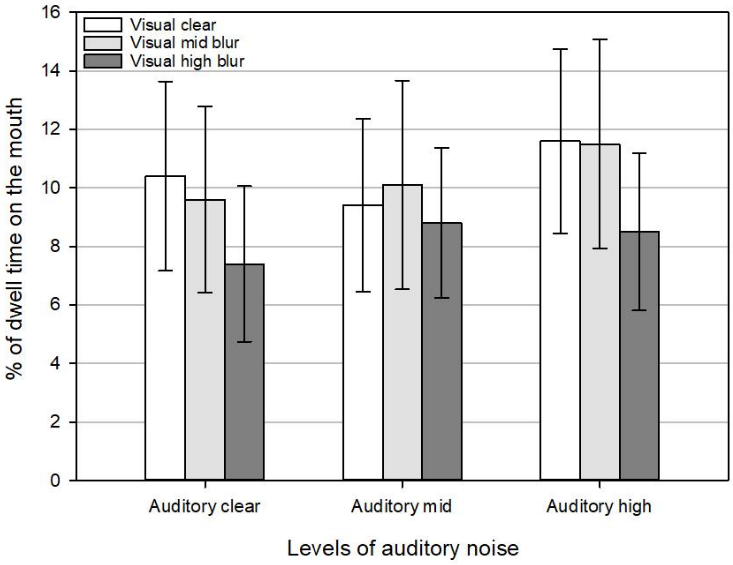


Figure 6.7 Dwell time on the mouth in auditory noise and visual blur, error bars represent 95% confidence intervals.

Pairwise comparisons for visual blur showed that there was no difference in dwell time in the clear condition compared to the mid blur condition (*p*= 1.00). The participants looked at the mouth significantly more in the clear condition compared to high blur (*p*= .036) and significantly more in the mid blur compared to high blur condition (*p*=.003). For auditory noise there was no difference in dwell time between the clear condition and the mid noise condition (*p*=1.00). The participants looked at the mouth significantly more in the high noise condition compared to the clear condition (*p*=.015) and significantly more in high noise compared to mid noise (*p*=.019).

6.3.5 Aim c: Do eye movements differ according to individual differences in visual benefit? (*N* = 42)

To investigate how eye movements differ according to individual differences in visual benefit the relationship between dwell time on the mouth and visual benefit was assessed in both AV and VO contexts. First the AV visual clear auditory high condition was used as this was the condition in which the participants received the most visual benefit. A Pearson correlation compared mouth dwell time in the visual clear auditory high noise condition with visual benefit however there was no significant correlation between the two variables *r* = -.020, *N*=42, *p* =.901. Separate correlations were conducted for each AOI to assess the relationship between dwell time and visual benefit on clear VO trials only as this is where the participants received the most visual benefit. There was no significant relationship between visual benefit (*M*=29%, *SD*=15%), and dwell time on the Mouth (*M*=20%, *SD*=14%), *r* = .190, *N*=42, *p* =.254; or eyes (*M*=11%, *SD*=10%), *r* = -.165, *N*=42, *p* =321. However, there was a significant weak positive relationship between visual benefit and fixating the nose (*M*=8%, *SD*=5%), *r* = .328, *N*=42, *p* =.044 suggesting that higher visual benefit scores were related to fixating the nose of a talker for longer.

6.3.6 Aim d: Is there a relationship between McGurk perception and visual benefit for words in noise? (*N* = 40)

McGurk responses were coded according to both fusion responses; ‘DA/THA’ and, non-auditory responses which include fusion responses and visual /ga:/ responses. For Talker 2 the congruent syllable GA was mistaken for DA for some people. Perception of the McGurk effect ranged from 0-100% across the participants for both fusion (*M*= 53%, *SD*= 31%) and non-auditory responses (*M*=71%, *SD*=27%).

Responses to congruent syllables were at ceiling for the majority of the participants, Table 6.1 reports the average percentage of correct responses to congruent stimuli. Responses to the GA stimulus spoken by talker 2 were low in comparison to other congruent stimuli as some of the participants confused this stimulus with DA.

Table 6.1

Means and standard deviations for congruent syllables for each talker

	Syllable	Mean	SD
Talker 1	/ba:/	88%	25%
	/ga:/	88%	30%
	/da:/	86%	21%
Talker 2			
	/ba:/	91%	20%
	/ga:/	57%	35%
	/da:/	89%	21%

A Pearson correlation was conducted to examine the relationship between McGurk perception (*M*= 70.6%, *SD*= 27%) and visual benefit (*M*= 29%, *SD*=15%). The following analyses used visual benefit scores from the high auditory noise with clear visual condition only, as this is the condition in which the participants benefited the most from seeing a face. There was no significant correlation between the two variables, *r* = -.075, *N*=40, *p* =.652. A Pearson correlation was also conducted to see if there was a relationship between visual benefit (high auditory noise, clear visual) and the congruent syllables, there was no significant relationship between visual benefit and each of the congruent syllable types: BA *r* = .139, *N*=40, *p* = .39, GA *r* = .019, *N*=40, *p* =.90, and DA *r* = .139, *N*=40, *p* =.39.

6.4 Discussion

The primary aims of this experiment were to explore how much benefit people received in different levels of auditory noise and visual blur, to explore where people looked in different levels of noise and blur, to explore whether people who looked at the mouth more gained more visual benefit, and to examine whether there was a relationship between visual benefit for word stimuli and McGurk perception.

The hypothesis that visual benefit and RT gain will increase as auditory noise increases and decrease as visual blur increases was partially supported (hypothesis a). For accuracy, there was only visual benefit when the auditory signal was in a high level of noise. Then, more benefit was received the clearer the visual signal was. Additionally, the participants only received RT gain when the visual signal was clear, in the mid auditory and high auditory noise conditions only. Overall the participants fixated the mouth more as auditory noise increased and less at the mouth as visual blurring increased (hypothesis b). It was expected that individuals with higher visual benefit would look at the mouth more (hypothesis c), however the nose was fixated on more as visual benefit increased. It was hypothesised that individuals who received higher visual benefit for words in noise would also be strong perceivers of the McGurk effect (hypothesis d, however, there was no significant relationship between McGurk perception and visual benefit.

6.4.1 Visual benefit & RT gain

Visual blur and auditory noise affected the amount of visual benefit people received. Generally, visual benefits were small and were only found when the auditory signal was degraded, and visual benefit for reaction time was only found when the visual signal was clear. Previous studies have found that visual gain does not follow the principle of inverse effectiveness (PoIE) which posits that AV integration increases in noise compared to no noise. The present study found evidence for the PoIE as visual benefit was only apparent in the highest level of auditory noise -20 dB. The levels of auditory noise used in the current experiment (-8dB, -20dB) were similar to that of Altieri and Wenger (2016) who found that visual gain was optimum at -18dB. The present experiment also used the addition of visual blur and as expected, visual benefit decreased as visual blur increased suggesting that the more degraded the visual stimulus is the less benefit there was in seeing a talker's face.

Altieri and Wenger (2013) calculated RT gain (AO-AV trials) and found that RT gain was greatest in the highest level of noise (-18 dB) meaning that the participants were much faster in difficult listening conditions when they could see the talker's face. This finding is supported by the present findings as RT was faster in clear and high (-20 dB) auditory noise. RT gain was also observed in the visual clear condition as RT was faster compared to the visual blur conditions meaning that seeing the talker's face was only beneficial when it was not visually degraded.

6.4.2 Eye movements

Consistent with Experiments 2 and 3, the participants in Experiment 4 looked at the mouth more when the visual signal was clear. Additionally they looked at the mouth more when there was a high level of auditory noise and less as visual blur increased suggesting that mouth gaze is dependent on how much benefit can be gained. Gaze behaviour also differed depending on the modality speech was presented in with the participants looking more at the eyes on AV trials but more at the mouth on VO trials. This suggests that the participants' strategies change depending on the information available to them. Fixating the mouth may be more beneficial in the visual only condition as that is the best strategy for gleaning speech information, whereas AV conditions may be more akin to natural conversation and so prompt looking at the eyes for social information.

Eye movements were examined in relation to visual benefit as it was expected that individuals with higher visual benefit would fixate the mouth more, consistent with Alsius et al (2016). However, there were no differences in eye movements according to visual benefit across all AOIs and on AV trials. The present study also aimed to extend the findings of Alsius et al. (2016) through including eye-tracking for the visual only condition. As visual benefit increased more time was spent looking at the nose on clear VO trials. There was no relationship between visual benefit and dwell time on the mouth or eyes in the visual clear condition. This suggests that a central position was adopted to view the face which may provide the best access to speech information (Buchan et al., 2007; Paré & Munhall., 2008). As fixating the mouth may not be necessary for visual benefit, this suggests that there are other factors which contributed to the participants' visual benefit such as, the ability to extract visual information via lip-reading. This seems likely as there was a positive relationship between visual benefit and lip-reading in the high visual blur condition suggesting that the benefit gained from seeing the talker's face was attributed to a superior ability to extract speech information from the face. It should be noted that the lip-reading task in the present experiment is relatively easy compared to a larger set size or more complex speech stimuli (Altieri et al., 2011; Sumby & Pollack, 1954).

6.4.3 McGurk Perception

In conjunction with experiments 1-3 in this thesis and previous literature (e.g. Basu-Mallick et al., 2015) McGurk perception varied from 0-100% and more non-auditory responses were reported more than fusion responses. Van Engen et al. (2017) found that there was no relationship between visual gain on a speech in noise task (SPIN) and McGurk perception. As accuracy on SPIN tasks and McGurk perception are both used as measures of AV integration, it follows that a strong perceiver of the McGurk effect would also be more accurate at identifying speech in noise as the illusion is dependent on integrating the auditory and visual information. The present study aimed to see if there was a relationship between visual benefit when words were presented in noise and McGurk perception. However, there was no relationship between individuals' visual benefit and McGurk perception. Grant and Seitz (1998) were able to find correlations (medium effect size) between congruent stimuli and incongruent stimuli with a sample size of 41. Therefore, a post-hoc power analysis was carried out to determine if the correlational analyses in Van Engen et al's (2017) study (*N*=38) and in the present study (*N*=40) were underpowered. Power was specified at 0.8, with a medium effect size (r = 0.5), this determined that a minimum sample size of 28 is required. As both studies appear to have sufficient power, further research is required to understand if the same mechanisms underpin McGurk perception and visual benefit to understand if these measures share the same mechanism for AV integration.

Several limitations of the present experiment should be noted. The dual task nature of the method may have increased cognitive load as participants were required to identify words in difficult listening situations as well as remembering the corresponding numbers on the keyboard for each word. This may have increased the task difficulty and slowed RT but was necessary to preserve eye-movement data. The inclusion of a learning block helped participants to memorise the key placement and this is reflected in the accuracy data which showed that participants were able to successfully complete the task.

Altieri and Wenger (2013) highlight a potential limitation of using a small set size of 8 words as using minimal words lacks ecological validity and may not reflect speech processing in real word contexts. However, the choice to minimise response options was taken in order to make comparisons between the present study and previous literature. Smaller set sizes also have the advantage of preserving shorter RT, which increases as response options increase (Hick, 1952). Wifall, Hazeltine and Mordkoff (2016) found that RT was slower (149ms slower) for 8 options compared to a 2 alternative force choice task, suggesting that whilst increasing the amount of responses from 2 to 8 increases RT the difference in time is minimal therefore 8 options were deemed appropriate.

6.4.4 Are words and phonemes really comparable?

The relationship between visual benefit for words in noise and the McGurk effect was examined, this was motivated by conflicting evidence in the literature regarding the relationship between speech in noise tasks and McGurk perception. Establishing whether or not this relationship exists would confirm if susceptibility to the McGurk effect and the ability to understand degraded speech share the same AV integration mechanism. One study found that the McGurk effect was not correlated with visual benefit for sentences in noise (Van Engen et al., 2017) whereas Grant and Seitz (1998) found that McGurk perception was correlated with the following speech stimuli presented in noise: nonsense syllables, consonants, and sentences. However, there was no relationship between McGurk perception and words identified in sentences (Grant & Seitz, 1998). This finding could be explained by the task differences employed in the study as the McGurk perception task was forced choice whereas an open-set task was used for the words. In the present work there are several differences between the McGurk task and the words task namely the noise manipulation was only present for the word stimuli, and the word task included eight response options whereas the McGurk task used four response options (mapped onto three keys) and increasing the amount of response options may also increase the task difficulty. Moreover, words may also require additional processes for lexical access and retrieval, and offer richer contextual cues compared to McGurk syllables. This may make it difficult to compare the tasks, therefore further analysis was conducted to see if there was a relationship between congruent syllables and words. As there was also no relationship between congruent syllables and congruent words, it is unclear whether the design of the present study was not sufficient to identify the relationship between different speech types or whether the relationship simply does not exist. Further research is required to assess the relationship between the McGurk effect and other measures of AV integration using appropriately matched tasks, this would establish what the McGurk effect is an appropriate measure of AV integration.

In sum, this experiment clarifies how visual benefit changes when stimuli are degraded and how eye movements are used to obtain visual information from the face of the talker. visual benefit increased as the clarity of the visual stimulus increased and the clarity of the auditory signal decreased, consistent with previous findings. RT gain was limited and only occurred when the visual stimulus was clear. There was no relationship between amount of visual benefit received when words were presented in noise and frequency of the McGurk effect.

Chapter 7: General Discussion

This chapter will draw comparisons between results from the four experiments reported in the thesis and discuss the implications of the findings. Ideas for future research will also be outlined.

7.1 Summary of results

The main findings from Experiments 1-4 were:

Experiment 1: Using a forced choice task increases instances of the McGurk effect compared to an open-set task. Different individuals perceive the McGurk effect to different extents, and different talkers produce illusionary percepts more reliably than others. Using a combination of the forced choice task and the fusion definition of the McGurk effect resulted in more instances of the illusion compared to the open-set task, whereas with the non-auditory definition fewer illusions were reported for the forced-choice task compared to the open-set task. Participants were faster to respond and more confident of their responses for congruent stimuli compared to incongruent stimuli.

Experiment 2: McGurk perception increased as auditory noise increased and decreased as visual blurring increased. In the highest level of visual blurring the McGurk effect was still perceived despite the highly degraded visual information. There was no relationship between time spent fixating the mouth and how often an individual perceived the McGurk effect. However, when participants perceived the McGurk effect they fixated the mouth longer compared to when they did not perceive the McGurk effect.

Experiment 3: This experiment replicated Experiment 2 with the addition that Vocoded speech was used as well as white noise to simulate the degraded speech experienced by cochlear implant users. The McGurk effect and dwell time on the mouth decreased as visual blur increased. There were no significant effects of auditory noise on eye movements or frequency of the McGurk effect.

Experiment 4: A different measure of AV integration was used (visual benefit). Visual benefit for RT and accuracy was greatest when the visual stimulus was clear. There was no relationship between McGurk perception and visual benefit for words presented in noise.

7.2 General Discussion

7.2.1 Exploring the McGurk effect as a measure of AV integration (Aim 1)

The first aim of the thesis was to explore the McGurk effect as a measure of AV integration. This was achieved in two ways firstly, through manipulating different methodological factors to see how this influenced perception of the McGurk effect and secondly, by comparing congruent speech stimuli with incongruent speech stimuli.

7.2.1.1 Individual differences in McGurk perception within talkers and participants

There was consistent variability in McGurk perception across all experiments and McGurk perception varied according to different talkers. Although McGurk perception differed substantially across individuals as shown in Table 7.1, average McGurk perception was similar across all experiments suggesting that the stimuli consistently elicited the illusion.

Table 7.1

Means and standard deviations and range for the McGurk effect across participants and experiments

Experiment	McGurk perception (Non-auditory)	
	Mean (SD)	Range
1	57% (21%)	19-98%
2	61% (9%)	25-78%
3	73% (9%)	55-92%
4	71% (27%)	0-100%
5	84% (22%)	67-100%

Although the same participants completed Experiments 2 and 3 McGurk perception increased in Experiment 3. The higher percentage of McGurk perception reported in Experiment 3 is most likely because of the auditory noise manipulation (vocoding + white noise) which would result in more visual responses. These results are in line with previous research (Basu Mallick, Magnotti & Beauchamp, 2015; Nath & Beauchamp, 2012) and confirm that there is substantial variability across individuals in how often they perceive the McGurk effect. Therefore, why individuals vary in how often they perceive the McGurk effect should be the focus of future research. The variability within participants and talkers makes it difficult to draw comparisons across the literature as every study uses their own participants and stimuli. Standardised instructions could be developed for researchers wishing to use the McGurk effect to reduce variability across stimuli (Alsius et al., 2017). Whilst there is significant variability in perception of the McGurk effect, individuals also differed substantially in the benefit they received from congruent AV speech perception (Experiment 4), therefore using congruent stimuli may not resolve the issue of variability.

7.2.1.2 The relationship between the McGurk effect and visual benefit

The thesis focused on one main measure of AV integration, the McGurk effect, the decision to use this particular measure was based on how prolific the illusion is in the literature. The McGurk effect is frequently reported as a robust illusion due to the ease with which it can be induced, it is also appealing as stimuli can be created easily. The research questions for the current thesis were developed in response to current literature. The thesis commenced in 2015, at which time the general consensus was positive regarding the usefulness of the McGurk effect. This was reflected in review papers such as Marques et al., (2016) which illustrated the varied applications of the McGurk effect. Subsequent review papers were published which criticised the McGurk effect (Alsius et al., 2017; Van Engen et al., 2017) as a valid measure of AV integration. Therefore, Experiment 4 was developed in response to criticisms of the McGurk effect and to specifically compare congruent AV speech with incongruent stimuli. It was expected that, in line with previous research (Grant & Seitz, 1998) there would be a relationship between different measures of AV integration and the McGurk effect. For example, strong perceivers of the McGurk effect would also be more accurate at identifying congruent speech in noise than weak perceivers of the McGurk effect because strong perceivers would be better at integrating information. The results of Experiment 4 showed that there was no relationship between visual benefit for words and the McGurk effect, in conjunction with previous findings (Van Engen et al., 2017). This suggests that care should be taken when when drawing conclusions directly by comparing the McGurk effect to AV integration during everyday conversation (Alsius et al., 2017; Van Engen et al., 2017). However, given the limited amount of research comparing the McGurk effect with other measures of AV integration further research is needed before definitive conclusions can be drawn regarding the McGurk effect as a measure of AV integration.

Until the nature of the McGurk effect can be established future research should adopt both congruent and incongruent speech measures for comparison. If future results were to clarify that the McGurk effect is categorically not a measure of AV integration the illusion still holds value for exploring related research questions including: 1) visual dominance/the benefit of visual information (discussed in Section 7.2.2 and 2) how incongruent multisensory information is resolved (discussed in Section 7.2.3).

7.2.2 AV integration and the benefit of visual information in quiet, and with degraded auditory and visual stimuli (Aim 2)

Two measures, visual benefit for words and McGurk perception were used to provide insights into AV integration and the influence of visual information. Auditory noise and visual blur were used to degrade the stimuli and thus manipulate the clarity of information from each modality. The experiments were able to establish how visual information is used in these contexts. Results will be discussed according to the different stimulus types.

7.2.2.1 Measuring the benefit of visual information to speech intelligibility

In Experiment 4 words were presented in visual only, auditory only or AV conditions. In quiet, there was no advantage with the addition of visual information as accuracy on the AV trials was the equivalent to the auditory only trials. RT was also faster on auditory only trials compared to AV trials. This suggests that visual information is redundant when the task is less demanding (in quiet) as auditory information alone is sufficient in quiet listening conditions.

The present findings also showed that visual benefit was only apparent in the highest level of auditory noise indicating that performance increased in the AV condition compared to the auditory only condition. This is in agreement with previous research which presented single syllable words in noise (de la Vaux &

Massaro, 2004). Overall, this suggests that visual information may be of most benefit when speech is presented in noise.

The clarity of the visual information was manipulated to see how much degradation can be tolerated whilst still inferring an advantage of AV speech compared to AO speech. Visual benefit was still observed even when the visual stimulus was severely degraded in the highest level of blur, although visual benefit increased as the clarity of the visual stimulus increased. Previous research shows that there is an advantage of seeing a talker's face even when the visual information is degraded (Brooke & Templeton, 1990; Jaekl et al., 2015; Jordan & Sergeant, 2000; Wozniak & Jackson, 1979). The specific degraded visual information used in the present research blurred the detail on the face so that only the key features of the face were visible. Performance was best when the visual information was clear, in line with previous research, which used similar visual degradation (Munhall et al., 2004; Tye-Murray, Spehar, Myerson et al., 2016). Thomas and Jordan (2002) used Gaussian blurring and found that whilst the ability to identify words decreased as visual blur increased, performance was still better for AV congruent words compared to auditory only words suggesting that there is still an advantage of seeing the talker's face even if the clarity of the visual stimulus is reduced.

The present findings add to this body of literature demonstrating the benefit of visual information in noise. Global movements of the face are enough to confer an advantage, and reduced spatial frequency information is sufficient for speech perception.

7.2.2.2 The McGurk effect in quiet listening conditions

McGurk syllables were presented in quiet and two levels of auditory noise. It was found that McGurk perception increased as auditory noise increased. This supports Sekiyama and Tokhura (1991) who found that the (Japanese) McGurk effect was not perceived in quiet but McGurk perception increased as auditory noise increased. Sekiyama et al. (2014) compared older adults with younger adults and calculated visual benefit using the accuracy scores (congruent trials – McGurk trials). They found that visual benefit increased as auditory noise increased. This suggests that when the auditory information is less reliable individuals focus more on the visual information, as the McGurk effect relies on the visual information this resulted in more illusory percepts.

McGurk perception increased as the clarity of the visual information increased and decreased the more the visual stimulus was blurred. Interestingly, the McGurk effect was still perceived in the highest level of visual blur in Experiment 2. Several experiments have found similar results, MacDonald et al. (2000) pixellated faces and found that when faces were severely pixilated the McGurk effect was still perceived. Thomas and Jordan (2002) used Gaussian blurring and found that whilst McGurk perception decreased with increased blurring McGurk perception was still observed for severely blurred faces.

Taken together, the results suggest that visual information is used more in auditory noise than quiet for both incongruent (McGurk) and congruent (words) AV speech, and whilst clear visual information has the most benefit for speech perception, fine detail in the features of the face are not necessary to benefit from visual speech information. These results demonstrate how the McGurk effect can be useful in establishing the weighting of auditory and visual information and the influence of visual information on auditory speech.

7.2.3 Eye movements in quiet and with degraded stimuli

The findings outlined in the previous section explain the importance of visual information. The goal of the eye-tracking experiments was to clarify how eye movements differ in background noise and using degraded visual stimuli, as where people look on a face may determine the quality of visual information they receive which could influence AV integration.

In all three eye tracking experiments the mouth was fixated the most followed by the nose, then the eyes. This pattern is in line with previous research (Alsius et al., 2016; Wilson et al., 2016). In Experiments 2 and 3 the hair and forehead were included as well as the chin/cheeks, these AOIs were fixated the least and therefore excluded from Experiment 4. Dwell time on the mouth was similar in all experiments as participants fixated the mouth on average 27%, 31%, and 20% of the time in Experiments 2, 3 and 4 respectively. Overall, the time spent looking at the mouth was lower than expected, as the mouth produces speech cues and provides important information such as, place of articulation (Amano & Sekiyama, 1998;

Summerfield, 1992) we would expect this area to be fixated the most. Alsius et al. (2016) found that when the visual stimulus was clear participants looked at the mouth >50% of the time for word stimuli. The differences in time spent looking at the mouth in the present experiments and previous research may be due to the particular talkers used in the current experiments compared to previous research. As Experiment 2 and 3 found Talker one's mouth was looked at more this suggests that some talkers may articulate more clearly and therefore provide better quality speech information, which would make fixating the mouth more advantageous.

7.2.3.1 Eye movements in noise

The same patterns in eye movements were also observed across experiments 2-4 for the different levels of visual blurring. Degrading the visual information had the effect that participants looked at the mouth less suggesting that looking at the mouth is only beneficial when the visual information is clear (Alsius et al., 2016; Paré, et al., 2003; Wilson et al., 2016). Whilst it was expected that participants would look more at the mouth of incongruent stimuli when the auditory signal was degraded this was not the case in Experiments 2 and 3. However, in Experiment 4 when congruent AV words were presented in high auditory noise participants looked at the mouth more compared to quiet. Previous research has suggested that where people look does not influence speech perception (Yi et al., 2013). However, other studies show that when stimuli are presented in auditory noise individuals looked more at the eyes for sentences (Vatikiotis-Bateson et al., 1998), more at the nose for incongruent stimuli (Paré, et al., 2003), and more at the nose and mouth when asked to identify key words in sentences (Buchan et al., 2008). In the present experiments, the second most fixated AOI was the nose which means participants could have also viewed the mouth peripherally. This suggests that adopting a more central view of the face may be preferable as this provides access to global speech information.

Overall, the findings suggest that when visual information is degraded gaze patterns are similar for congruent and incongruent speech as looking at the mouth decreases as blur increases. In auditory noise, looking at the mouth may only be useful for congruent speech stimuli compared to congruent stimuli as the auditory and visual modalities confer complementary speech cues.

7.2.3.2 The relationship between eye movements and AV integration

The relationship between fixating the mouth and AV integration is presently unclear in particular, how important fixating a talker's mouth is for AV integration. In line with previous research (Paré, et al., 2003), it was found that there was no relationship between the time individuals spent looking at the mouth of incongruent stimuli and how often they perceived the McGurk effect, in quiet. This suggests that looking at the mouth is not necessary to perceive the McGurk effect. Instead, when people perceive the McGurk effect they may be able to extract speech information from other parts of the face or make use of peripheral vision. However, when dwell time on the mouth was collapsed across all conditions, overall participants spent longer looking at the mouth when they perceived the McGurk effect compared to when they did not perceive the McGurk effect. For example, participant 16 looked at the mouth 4% of the time in quiet but perceived the McGurk effect 61% of the time, whereas when mouth dwell time was collapsed across noise conditions, this participant looked at the mouth 41% when they did not perceive the McGurk effect and 66% when they did perceive the McGurk effect. This suggests that, overall increased AV integration is accompanied by increased time spent looking at the mouth. These results show that the McGurk effect was useful in elucidating what part of the visual speech information is important when auditory and visual information is incongruent.

There was also no relationship between time spent looking at the mouth and visual benefit in quiet listening conditions. Theses findings support research which finds that fixating the mouth did not influence AV integration for congruent speech (sentences; Everdell et al., 2007). This suggests that participants who experience more visual benefit are better at extracting visual information from the face as a whole.

7.2.3 The timing of AV integration (Aim 3)

The third aim was to examine the temporal properties of AV integration to inform AV integration theory, this was explored in Experiment 1 and 4 with behavioural measures (RT). Experiment 1 found that RT was slower for incongruent stimuli compared to congruent stimuli. The results supported those of Massaro and Cohen (1983) who hypothesised that longer RT for incongruent stimuli compared to congruent stimuli was indicative of the time taken to process stimuli. Taken together, these findings may be indicative of the decision making process outlined in the FLMP (Massaro & Oden, 1980). This theory suggests that speech perception involves identifying and comparing the acoustic properties of the incoming signal with prototypes held in memory (feature evaluation, and prototype matching), the features of the stimulus are matched with the most appropriate prototype and the stimulus is identified (pattern classification).

In Experiment 4 RT was faster for AV words than auditory only words when the visual signal was clear only. Altieri and Wenger (2013) found that RT was faster (~700ms) in the AV condition than the AO condition (Altieri & Townsend, 2011). Faster RT for congruent AV speech compared to AO in this study (Altieri & Townsend, 2011) and the present research suggests an advantage of AV information over AO and that the addition of seeing the talker's face facilitates faster speech processing.

These results show that there are differences in RT depending on whether speech is congruent or incongruent. For congruent stimuli RT was faster compared to incongruent stimuli (Experiment 1), RT was also faster for congruent AV stimuli compared to AO stimuli (Experiment 4). This may reflect the decision making process - when speech is congruent the addition of visual information is beneficial for understanding speech and therefore auditory and visual information are integrated earlier resulting in faster responses. In the case of incongruent stimuli, the auditory and visual information provide conflicting information which must be resolved by either opting for the auditory or visual modality or combining both resulting in an illusory percept, these additional decisions may be reflected in the longer RT (Massaro & Cohen, 1983). Therefore, the McGurk effect can be considered a measure of the speed of conflict resolution. The finding that RT was longer for incongruent stimuli regardless of whether or not an illusion was perceived suggests that this decision making process always takes place in relation to incongruent stimuli.

7.3 Potential Implications of findings

Through providing knowledge of the benefit of visual information, the findings have implications for a) theoretical accounts of speech perception, b) individuals with hearing impairments, c) understanding AV integration in general and, d) methodology used in future research.

Evidence for the FLMP (Massaro & Oden, 1980) was found in Experiment 1 as RT was faster for congruent speech compared to incongruent speech suggesting that incongruent speech is resolved later than congruent speech. The finding from Experiment 4 that RT was faster for AV speech compared to AO speech would seem to suggest that congruent AV information is combined earlier however this conclusion is tentative without evidence from ERPs to corroborate these findings. Overall, the findings suggest that AV congruent speech is processed earlier in time compared to incongruent speech or AO speech.

Previous literature suggested that looking at the mouth was not necessary for McGurk perception which the thesis confirmed as there was no relationship between time spent looking at the mouth and McGurk perception, however the results also suggest that overall, when dwell time is included across all noise conditions, looking at the mouth increases the likelihood of McGurk perception. This suggests that looking at the mouth of a talker would still be beneficial for speech perception in noise. This finding could be used to conduct further research with individuals with HI with the aim of improving AV integration in noise.

The findings can be interpreted in the context of AV integration in general and how information from each sense is weighted. The modality appropriate hypothesis posits that the most reliable sense dominates (Welch & Warren, 1980). This is supported by the findings using congruent AV speech, as visual benefit was observed in high auditory noise only suggesting that when the auditory signal was less reliable the visual stimulus dominated, for incongruent speech vision influenced audition in all noise contexts. The findings showed that when the auditory signal was clear and the visual information was blurred the McGurk effect was still perceived suggesting that the the visual information provided by the face is highly robust to visual blurring.

Several methodological factors were explored in the thesis which could have implications for researchers wishing to use the McGurk effect as a measure of AV integration. If researchers wish to examine the McGurk effect, then a paradigm which elicits the most amount of McGurk responses is desirable therefore, a forced choice task should be used. Fixation cross location was also manipulated in Experiments 2 and 3 and was found to influence dwell time on the mouth, therefore a peripheral fixation cross is preferable in future eye-movement experiments.

7.4 Methodological strengths and limitations

The thesis used different methods including behavioural and eye-tracking, to answer the research questions. Whilst previous literature has mostly been concerned with understanding speech in auditory noise, the benefit of visual information is not well understood. Understanding this is important for everyday communication, and quality of life for people with hearing or visual impairments. The thesis aimed to address this by understanding how eye movements are related to AV integration and how auditory and visual information interact. Few studies (Alsius et al., 2016; McGettigan et al., 2012; Munhall et al., 2004; Tye-Murray, Spehar, Myerson et al., 2016) have simultaneously manipulated the quality of the auditory and visual information. The findings that McGurk effect perception varied across individuals, and the pattern of eye movements in noise were consistent across experiments and were able to contribute to understanding in the field. A further strength is that the stimuli and method used were piloted extensively in the first experiment, and comparisons between talkers were used throughout.

There are several limitations of the thesis as a whole which should be discussed. This thesis focused on the McGurk effect to gain insights into AV integration. A potential limitation of the present experiments is that one type of McGurk syllable (A_{BA} V_{GA}) was used per talker. The stimulus $A_{BA}V_{GA}$ is also the most well known and most widely used, due to previous reports that this particular stimulus type produces the illusion the most frequently (McGurk & MacDonald, 1976). Furthermore, Amano and Sekiyama (1998) reported that a smaller stimulus set-size results in increased instances of the McGurk effect compared to a large stimulus set-size. The particular talkers used in the current experiments were based on the results from Experiment 1 as stimuli which produced the McGurk effect to the greatest extent were used. However, different participants may perceive the McGurk effect to different extents based on the particular stimulus used (Basu Mallick et al., 2015). Therefore, the results may have been influenced by the choice of stimuli used in the current experiments. Despite this, the same pattern of behavioural and eye movement results observed in the present experiments were also observed in several other studies (Alsius et al., 2016; Basu Mallick et al., 2015) which used different stimuli.

A potential issue also concerns the definition of the McGurk effect used within the thesis. Coding McGurk responses as anything other than the auditory signal means that errors caused by fatigue or inattention could be counted as McGurk responses. However, the findings from Experiments 2 and 3 show that McGurk responses were systematically affected by the manipulations of auditory noise and visual blur, which suggests that any such errors are likely to be minimal and have little influence on the overall pattern of results or were averaged out between conditions.

Limitations of eye movement measures should be acknowledged. The use of eye tracking in a laboratory context may not recreate natural gaze behaviour as speech is considered social and gaze helps to facilitate this during naturalistic conversation. As such, it may be that the nature of the study prevented any observable differences in eye movements according to the amount of visual benefit people received in Experiment 4. The fixation behaviour reported in the current experiments may differ from that during conversation, for example, viewers may look more at the eyes of a talker during conversation for social cues (Itier & Batty, 2009). Therefore, focusing on the eyes may be more useful for longer speech stimuli such as sentences whereas the present study used short stimuli (~2000ms). Future research could build on the present findings by using longer speech stimuli e.g. sentences in comparison with the McGurk effect. Previous findings (Buchan et al., 2008) also suggest that talker identity can influence gaze, as when a different talker is presented on every trial, participants focus more on the mouth compared to when the talker was consistent across trials. As talker identity may have influenced time spent fixating the mouth the four talkers were compared. Although talker identity was randomised across trials, it was found that participants tended to fixate more on the mouth of one particular talker, this suggests that the way in which this talker articulated the syllables provided more speech cues, therefore it was more beneficial to look at the mouth of this talker.   However, similar patterns in eye movements have been identified across Experiments 2-4 which is promising in terms of building a reliable picture of which parts of the visual stimulus are important.

Eye-tracking experiments involve the subjective creation of AOIs which can be placed on the features of a face to determine where people look. These often differ in the shape, size and quantity used across studies. This can make comparisons with the current experiments and across different studies difficult. It is also not clear if a fixation equates to attention on that area or if peripheral vision is being utilised. Experiments 2 and 3 used 6 AOIs whereas Experiment 4 used 3 AOIs. The decision was taken to minimise the number of AOIs in Experiment 4 to streamline analysis and because Experiments 2 and 3 showed that people did not look at AOIs which included the hair as often as key features of the face such as nose, eyes and mouth, therefore these were removed in Experiment 4.

Overall, the aims of the thesis were quite broad and arose from the controversies identified in the literature, exploration of these aims allowed greater breadth of knowledge. Alternatively, having more streamlined aims would have enabled more in depth examination of different types of speech, for example, a wider variety of stimuli could have been used including sentence stimuli, which would have allowed for greater comparisons and provided more insight into the McGurk effect as a measure of AV integration.

7.5 Future research

All experiments in the thesis used a specific population of young adult NH listeners. Therefore, the results may only pertain to adults in this particular age range (18-48 years old). Young adults with NH were used as there are often cognitive deficits associated with older adults and individuals with hearing impairments. The ability to integrate auditory and visual information varies across older adults (Sekiyama et al., 2014) and people with hearing impairments (Tye-Murray, Spehar, Sommers et al., 2016) therefore future research should include these populations to gain a full understanding of how AV integration varies across individuals. Conducting experiments with NH listeners is also advantageous as variables can be manipulated which would be difficult with hearing impaired listeners. For example, the parameters of the vocoder used in Experiment 3 were easily manipulated.

7.5.1 AV integration across the life-span

Research shows that AV integration can change across the life span, and that individuals of different ages may differ in their susceptibility to visual information, for example, children appear to be more attuned to auditory information and experience the McGurk effect less often than adults (Tremblay et al., 2007). An experiment could be conducted using a similar method to Experiment 2 in which visual and auditory information is degraded and McGurk perception compared across children and adults. Understanding how children are influenced by visual information could have implications for how children learn and interact in the classroom.

Older adults experienced more visual benefit compared to younger adults (Sekiyama et al., 2014). Examining how older adults integrate auditory and visual information in noisy situations is important given the prevalence of auditory and visual impairments experienced by older adults. Understanding how older adults use AV information is also of particular importance as it can help to improve cognitive deficits associated with ageing and therefore, may improve older adults' quality of life. For example, Peiffer, Mozolic, Hugenschmidt, and Laurienti (2007) found that older adults are faster than younger adults to respond when stimuli are AV compared to auditory or visual only.

7.5.2 Individuals with hearing impairments

Individuals with HI may use visual information differently from adults with NH, for example if they have been deaf since childhood they may have learnt to rely more on visual information (UK cochlear implant group, 2004). Due to the large variability in the success of cochlear-implants, understanding how hearing impaired listeners use visual information when listening in noise is of interest. Studies aimed at training CI users have been successful to different degrees (Henshaw & Ferguson, 2013), therefore understanding factors which can contribute to improving the user's experience with CIs is important so that they get the best out of hearing. Future research could use a paradigm similar to the one presented in Experiment 2 to examine AV integration with individuals with hearing impairments.

Given the results of Experiment 4 multiple measures of AV integration including incongruent and congruent speech should be used in future experiments which include the McGurk effect.

The use of Electroencephology (EEG) to examine the timing of AV integration

The temporal properties of AV integration could also be explored by using neuroimaging techniques in conjunction with RT. The use of RT in the present experiments is limited as it is unclear whether RT reflects the temporal properties of AV integration or the decision making process. Investigating RT and related cortical activity (event-related potentials) in response to AV incongruent stimuli could help us to understand the mechanisms behind perception of congruent and incongruent speech. As discussed in Chapter 1, whether auditory and visual information converge earlier or later in time has been debated, therefore, further research is required to establish at what point auditory and visual information converges.

7.6 Original Contribution to Knowledge

Previous literature has established that seeing a talker's face aids speech perception. What is unclear is how visual information is used and what part of the visual information is important. Understanding how visual information can benefit communication is important for NH listeners as well as hearing impaired listeners. The experiments conducted explored the influence of visual information when speech is degraded, and how visual information is gathered from eye movements. The mechanisms behind AV integration and how auditory and visual information converges are not well understood. The thesis provides an original contribution to knowledge as the findings were able to resolve some inconsistences in the literature regarding whether looking at the mouth of a talker is important, these results have subsequently been submitted for publication (Stacey, Howard, Mitra & Stacey, under review). Furthermore, evidence for theories of speech perception was provided. An exploration of different measures of AV integration allowed recommendations to be made as to how the field can move forward.

7.7 Conclusion

The findings were able to elucidate the role of visual information when both the auditory and visual information are degraded. In quiet, looking at the mouth of the talker is not necessary for AV integration. In auditory noise, visual information is of most benefit when it is clear but fine detail on the face is not needed for AV integration. In noise, more time is spent looking at the mouth for congruent stimuli only. As visual blur increases looking at the mouth decreases for both congruent and incongruent stimuli. Incongruent AV speech may be integrated later compared to congruent AV speech. The validity of the McGurk effect as a measure of AV integration should be explored in future research; however, the current findings demonstrated the usefulness of the illusion for exploring the influence of visual information on auditory perception in quiet and noise.

Glossary

Abbreviations

A	Auditory
AO	Auditory only
ANOVA	Analysis of variance
AV	Audio-visual
CI	Cochlear implant
dB	Decibels
DFI	Double flash illusion
DRT	Direct realist theory
ERPs	Event-related-potentials
FLMP	Fuzzy logic motor perception
HI	Hearing impairment
NH	Normal hearing
PIOE	Principle of inverse effectiveness
RMT	Revised motor theory
SNR	Signal to noise ratios
SRT	Speech reception threshold
STS	Superior temporal sulcus
SPL	Sound pressure level
V	Visual
VO	Visual only

Phonetic notation

/a:/	Ah
/θa:/	Tha
/Ba:/	Ba
/Ga:/	Ga
/Da:/	Da
/bæt/	Bat
/væt/	Vat
/ Ɔ:/	Or

References

Akeroyd, M. A., Foreman, K., & Holman, J. A. (2014). Estimates of the number of adults in England, Wales, and Scotland with a hearing loss. *International Journal of Audiology, 53*(1), 60-61. doi:10.3109/14992027.2013.850539

Alsius, A., Möttönen, R., Sams, M. E., Soto-Faraco, S., & Tiippana, K. (2014). Effect of attentional load on audiovisual speech perception: Evidence from ERPs. *Frontiers in Psychology, 5,* 727. doi:10.3389/fpsyg.2014.00727

Alsius, A., Paré, M., & Munhall, K. G. (2017). Forty years after Hearing lips and seeing voices: The McGurk effect revisited. *Multisensory Research, 31*(1-2), 111-144. doi:10.1163/22134808-00002565

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, and Psychophysics, 78*(5), 1472-1487. doi:10.3758/s13414-016-1109-4

Altieri, N. & Wenger, M. J. (2013). Neural dynamics of audiovisual speech integration under variable listening conditions; an individual participant analysis. *Frontiers in Psychology, 4,* 1-15. doi:10.3389/fpsyg.2013.00615

Altieri, N., & Townsend, J. T. (2011). An assessment of behavioral dynamic information processing measures in audiovisual speech perception. *Frontiers in Psychology, 2,* 238. doi:10.3389/fpsyg.2011.00238

Amano, J., & Sekiyama, K. (1998). The McGurk effect is influenced by the stimulus set size. In *AVSP'98 International Conference on Auditory-Visual Speech Processing* (pp.43-48). Retrieved from https://www.isca-speech.org/archive

Andersen, T. S., Tiippana, K., & Sams, M. (2004). Factors influencing audiovisual fission and fusion illusions. *Cognitive Brain Research*, *21*(3), 301-308. doi: 10.1016/j.cogbrainres.2004.06.004

Apoux, F., & Bacon, S. P. (2008). Differential contribution of envelope fluctuations across frequency to consonant identification in quiet. *The Journal of the Acoustical Society of America*, *123*(5), 2792-2800. doi:10.1121/1.2897916

Arizpe, J., Kravitz, D. J., Yovel, G., & Baker, C. I. (2012). Start position strongly influences fixation patterns during face processing: Difficulties with Eye Movements as a measure of information use. *PloS One*, *7*(2), 31106. doi:10.1371/journal.pone.0031106

Arnold, P., & Hill, F. (2001). Bisensory augmentation: A speechreading advantage when speech is clearly audible and intact. *British Journal of Psychology, 92*(2), 339-355. doi:10.1348/000712601162220

Auer, E. T., JR., Bernstein, L. E., Waldstein, R. S., & Tucker, P. E. (1997). Effects of phonetic variation and the structure of the lexicon on the uniqueness of words. In C. Benoit & R. Campbell (Eds.), *Proceedings of the ESCAIESCOP Workshop on Audio-Visual Speech Processing* (pp. 21-24). Rhodes, Greece, September 26-27.

Auer, E. T., JR., & 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. doi: 10.1044/1092-4388(2007/080)

Bastien-Toniazzo, M., Stroumza, A., & Cavé, C. (2010). Audio-visual perception and integration in developmental dyslexia: An exploratory study using the McGurk effect. *Behaviour, Brain and Cognition, 25*(3). Retrieved from http://journals.openedition.org/cpl/4928

Basu Mallick, D. B., Magnotti, J. F., & Beauchamp, M. S. (2015). Variability and stability in the McGurk effect: Contributions of participants, stimuli, time, and response type. *Psychonomic Bulletin and Review, 22*(5), 1299-1307. doi:10.3758/s13423-015-0817-4

Bear, H. L., & Harvey, R. (2017). Phoneme-to-viseme mappings: The good, the bad, and the ugly. *Speech Communication, 95*, 40-67. doi:10.1016/j.specom.2017.07.001

Beauchamp, M. S., Nath, A. R., & Pasalar, S. (2010). fMRI-Guided transcranial magnetic stimulation reveals that the superior temporal sulcus is a cortical locus of the McGurk effect. *Journal of Neuroscience, 30*(7), 2414-2417. doi:10.1523/JNEUROSCI.4865-09.2010

Benguerel, A. P., & Pichora-Fuller, M. K. (1982). Coarticulation effects in lipreading. *Journal of Speech, Language, and Hearing Research, 25*(4), 600-607. doi:10.1044/jshr.2504.600

Benoit, M. M., Raij, T., Lin, F. H., Jääskeläinen, I. P., & Stufflebeam, S. (2010). Primary and multisensory cortical activity is correlated with audiovisual percepts. *Human Brain Mapping, 31*(4), 526-538. doi:10.1002/hbm.20884.

Bernstein, L. E., & Liebenthal, E. (2014). Neural pathways for visual speech perception. *Frontiers in Neuroscience*, *8*, 386. doi:10.3389/fnins.2014.00386

Bernstein, L. E., Tucker, P. E., & Demorest, M. E. (2000). Speech perception without hearing. *Perception and Psychophysics*, *62*(2), 233-252. doi:10.3758/BF03205546

Blackburn, C. L. (2019). The benefit received from visual information when listening to clear and degraded speech in background noise. (Unpublished doctoral dissertation). Nottingham Trent University, Nottingham.

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*. doi:10.1177/2331216519837866

Brancazio, L. (2004). Lexical influences in audiovisual speech perception, Journal of Experimental Psychology: *Human Perception and Performance, 30*, 445–463. doi: https://doi.org/10.1037/0096-1523.30.3.445

Brancazio, L., & Miller, J. L. (2005). Use of visual information in speech perception: Evidence for a visual rate effect both with and without a McGurk effect. *Perception and Psychophysics*, *67*(5), 759-769. doi.org/10.3758/BF03193531

Brown, V. A., Hedayati, M., Zanger, A., Mayn, S., Ray, L., Dillman-Hasso, N., & Strand, J. F. (2018). What accounts for individual differences in susceptibility to the McGurk effect?. *PloS one, 13*(11). doi:10.1371/journal.pone.0207160

Brungart, D. S., Simpson, B. D., Ericson, M. A., & Scott, K. R. (2001). Informational and energetic masking effects in the perception of multiple simultaneous talkers. *The Journal of the Acoustical Society of America*, *110*(5), 2527-2538. doi:10.1121/1.1408946

Buchan, J. N., & Munhall, K. G. (2012). The effect of a concurrent cognitive load task and temporal offsets on the integration of auditory and visual speech information. *Seeing and Perceiving, 25*, 87–106. doi:10.1163/187847611x620937

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

Buchan, J. N., 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. doi:10.1016/j.brainres.2008.06.083

Campbell, C. S., & Massaro, D. W. (1997). Perception of Visible Speech: Influence of Spatial Quantization. *Perception, 26*(5), 627–644. doi:10.1068/p260627

Ciorba, A., Bianchini, C., Pelucchi, S., & Pastore, A. (2012). The impact of hearing loss on the quality of life of elderly adults. *Clinical Interventions in Aging*, *7*, 159. doi: 10.2147/CIA.S26059

Colin, C., Radeau, M., & Deltenre, P. (2005). Top-down and bottom-up modulation of audiovisual integration in speech. *European Journal of Cognitive Psychology*, *17*(4), 541-560. doi:10.1080/09541440440000168

Colin, C., Radeau, M., Soquet, A., Demolin, D., Colin, F., & Deltenre, P. (2002). Mismatch negativity evoked by the McGurk–MacDonald effect: A phonetic representation within short-term memory. *Clinical Neurophysiology*, *113*(4), 495-506. doi.org:10.1016/S1388-2457(02)00024-X

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. doi:10.1037/0096-3445.134.2.222

de la Vaux, S. K., and Massaro, D. W. (2004) Audiovisual speech gating: Examining information and information processing. *Cognitive Processing*, *5*, 106-112. doi:10.1007/s10339-004-0014-2

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. doi:10.1044/jshr.3504.876

Diehl, R. L. (1987). Auditory constraints on speech perception. In Schouten M.E.H. (Ed.) *The Psychophysics of Speech Perception* (pp.210-219). Dordrecht: Springer.

Dodd, B., Plant, G., & Gregory, M. (1989). Teaching lip-reading: The efficacy of lessons on video. *British Journal of Audiology*, *23*(3), 229-238. doi:10.3109/03005368909076504

Dorman, M. F., & Loizou, P. C. (1997). Speech intelligibility as a function of the number of channels of stimulation for normal-hearing listeners and patients with cochlear implants. *The American Journal of Otology*, *18*(6), 113-114. Retrieved from https://www.journals.elsevier.com/journal-of-otology/

Dorman, M. F., & Loizou, P. C. (1998). The identification of consonants and vowels by cochlear implant patients using a 6-channel continuous interleaved sampling processor and by normal-hearing subjects using simulations of processors with two to nine channels. *Ear and Hearing*, *19*(2), 162-166. doi:10.1097/00003446-199804000-00008

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. doi:10.1121/1.419603

Dorman, M. F., Loizou, P. C., Fitzke, J., & Tu, Z. (1998). The recognition of sentences in noise by normal-hearing listeners using simulations of cochlear-implant signal processors with 6–20 channels. *The Journal of the Acoustical Society of America, 104*(6), 3583-3585. doi:10.1121/1.423940

Dorman, M. F., Natale, S. C., Butts, A. M., Zeitler, D. M., & Carlson, M. L. (2017). The Sound Quality of Cochlear Implants: Studies with single-sided deaf Patients. *Otology and Neurotology, 38*(8), 268-273. doi:10.1097/MAO.0000000000001449.

Easton, R. D., & Basala, M. (1982). Perceptual dominance during lipreading. *Perception and Psychophysics, 32*, 562–570. doi:10.3758/BF03204211

Ernst, M. O., & Bülthoff, H. H. (2004). Merging the senses into a robust percept. *Trends in Cognitive Sciences*, *8*(4), 162-169. doi:10.1016/j.tics.2004.02.002

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

Faul, F., Erdfelder, E., Lang, A.-G. & Buchner, A. (2007). G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavioral Research Methods, 39,* 175–91. doi:10.3758/BF03193146

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*. doi:10.1044/1092-4388(2009/08-0137)

Findlay, J. M., & Gilchrist, I. D. (2003). Active vision: The psychology of looking and seeing. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198524793.001.0001

Fishman, K. E., Shannon, R. V., & Slattery, W. H. (1997). Speech recognition as a function of the number of electrodes used in the SPEAK cochlear implant speech processor. *Journal of Speech, Language, and Hearing Research*, *40*(5), 1201-1215. doi:10.1044/jslhr.4005.1201

Fixmer, E., & Hawkins, S. (1998). The influence of quality of information on the McGurk effect. In *AVSP'98 International Conference on Auditory-Visual Speech Processing*. Retrieved from https://www.isca-speech.org/archive

Fowler, C. A. (1981). Production and perception of coarticulation among stressed and unstressed vowels. *Journal of Speech, Language, and Hearing Research*, *24*(1), 127-139. doi:10.1044/jshr.2401.127

Fowler, C. A. (1996). Listeners do hear sounds, not tongues. *The Journal of the Acoustical Society of America*, *99*(3), 1730-1741. doi:10.1121/1.415237

Galantucci, B., Fowler, C. A., & Turvey, M. T. (2006). The motor theory of speech perception reviewed. *Psychonomic Bulletin and Review*, *13*(3), 361-377. doi:10.3758/BF03193990

Gantz, B. J., Woodworth, G. G., Knutson, J. F., Abbas, P. J., & Tyler, R. S. (1993). Multivariate Predictors of Audiological Success with Multichannel Cochlear

Implants. *Annals of Otology, Rhinology and Laryngology*, *102*(12), 909–916. doi:10.1177/000348949310201201

Gatehouse, S., & Gordon, J. (1990). Response times to speech stimuli as measures of benefit from amplification. *British Journal of Audiology, 24*(1), 63-68. doi:10.3109/03005369009077843

Giard, M. H., & Peronnet, F. (1999). Auditory-visual integration during multimodal object recognition in humans: a behavioral and electrophysiological study. *Journal of Cognitive Neuroscience, 11*(5), 473-490. doi:10.1162/089892999563544

Glasberg, B. R., & Moore, B. C. (1990). Derivation of auditory filter shapes from notched-noise data. *Hearing Research, 47(*1-2), 103-138. doi:10.1016/0378-5955(90)90170

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. doi: 10.1121/1.1928892

Grant, K. W., & Seitz, P. F. (1998). Measures of auditory–visual integration in nonsense syllables and sentences. *The Journal of the Acoustical Society of America*, *104*(4), 2438-2450. doi:10.1121/1.423751

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. doi:10.1121/1.1288668

Green, K. P., & Gerdman, A. (1995). Cross-modal discrepancies in coarticulation and the integration of speech information: The McGurk effect with mismatched vowels. *Journal of Experimental Psychology: Human Perception and Performance, 21*(6), 1409-1426. doi:10.1037/0096-1523.21.6.1409

Gurler, D., Doyle, N., Walker, E., Magnotti, J., & Beauchamp, M. (2015). A link between individual differences in multisensory speech perception and eye movements. *Attention, Perception, and Psychophysics, 77*(4), 1333-1341. doi:10.3758/s13414-014-0821-1

Halle, M., & Stevens, K. (1962). Speech recognition: A model and a program for research. *IRE Transactions on Information Theory, 8*(2), 155-159. doi:10.1109/TIT.1962.1057686

Hayhoe, M., & Ballard, D. (2005). Eye movements in natural behavior. *Trends in Cognitive Sciences, 9*(4), 188-194. doi:10.1016/j.tics.2005.02.009

Henshaw, H., & Ferguson, M. A. (2013). Efficacy of individual computer-based auditory training for people with hearing loss: A systematic review of the evidence. *PloS one, 8*(5), e62836. doi:10.1371/journal.pone.0062836

Hick, W. E. (1952). On the rate of gain of information. *Quarterly Journal of Experimental Psychology, 4,* 11–26. doi:10.1080/17470215208416600

Hirst, R. J., Stacey, J. E., Cragg, L., Stacey, P. C., & Allen, H. A. (2018). The threshold for the McGurk effect in audio-visual noise decreases with development. *Scientific Reports, 8*(1), 12372. doi:10.1038/s41598-018-30798-8

Hoffman, J. E., & Subramaniam, B. (1995). The role of visual attention in saccadic eye movements. *Perception and Psychophysics, 57*(6), 787-795. doi:10.3758/BF03206794

Howard, I. P., & Templeton, W. B. (1966). *Human spatial orientation*. Oxford, England: John Wiley & Sons.

Itier, R. J., & Batty, M. (2009). Neural bases of eye and gaze processing: The core of social cognition. *Neuroscience and Biobehavioral Reviews, 33*(6), 843-863. doi:10.1016/j.neubiorev.2009.02.004

Jaekl, P., Pesquita, A., Alsius, A., Munhall, K., & Soto-Faraco, S. (2015). The contribution of dynamic visual cues to audiovisual speech perception. *Neuropsychologia, 75,* 402-410. doi:10.1016/j.neuropsychologia.2015.06.025

Jesse, A., Vrignaud, N., Cohen, M. M., & Massaro, D. W. (2000). The processing of information from multiple sources in simultaneous interpreting. *Interpreting*, *5*(2), 95-115. doi:10.1075/intp.5.2.04jes

Jiang, J., & Bernstein, L. E. (2011). Psychophysics of the McGurk and other audiovisual speech integration effects. *Journal of Experimental Psychology: Human Perception and Performance, 37*(4), 1193-1209. doi:10.1037/a0023100

Jordan, T., & Sergeant, P. (2000). Effects of distance on visual and audiovisual speech recognition. *Language and Speech, 43,* 107–124. doi:10.1177/00238309000430010401

Kelly, M. C., & Chen, P. (2009). Development of form and function in the mammalian cochlea. *Current Opinion in Neurobiology*, *19*(4), 395-401. doi:10.1016/j.conb.2009.07.010

Laneau, J., Moonen, M., & Wouters, J. (2006). Factors affecting the use of noise-band vocoders as acoustic models for pitch perception in cochlear implants. *The Journal of the Acoustical Society of America*, *119*(1), 491-506. doi:10.1121/1.2133391

Lansing, C. R., & McConkie, G.W. (1999). Attention to facial regions in segmental and prosodic visual speech perception tasks. *Journal of Speech, Language, and Hearing Research, 42*, 526-539. doi:10.1044/jslhr.4203.526

Liberman, A. M. (1957). Some results of research on speech perception. *The Journal of the Acoustical Society of America*, *29*(1), 117-123. doi:10.1121/1.1908635

Liberman, A. M., & Mattingly, I. G. (1985). The motor theory of speech perception revised. *Cognition*, *21*(1), 1-36. doi:10.1016/0010-0277(85)90021-6

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. *Journal of American Medical Association: Internal Medicine, 173*(4), 293-299. doi:10.1001/jamainternmed.2013.1868

Lisker, L., & Abramson, A. S. (1964). A cross-language study of voicing in initial stops: Acoustical measurements. *Word, 20*(3), 384-422. doi:10.1080/00437956.1964.11659830

Loizou, P. C. (1998). Mimicking the human ear. *IEEE Signal Processing Magazine, 15*(5), 101-130. doi:10.1109/79.708543

Loizou, P. C., Dorman, M., & Tu, Z. (1999). On the number of channels needed to understand speech. *The Journal of the Acoustical Society of America, 106*(4), 2097-2103. doi:10.1121/1.427954

Lucey, P., Martin, T., & Sridharan, S. (2004). Confusability of phonemes grouped according to their viseme classes in noisy environments. In Cassidy, S, Cox, F, Mannell, R, & Palethorpe, S (Eds.) *Proceedings of the 10th Australian International Conference on Speech Science and Technology* (pp.265-270). Sydney, NSW: Australasian Speech Science and Technology Association. Retrieved from https://assta.org

Ma, W. J., Zhou, X., Ross, L. A., Foxe, J. J., & Parra, L. C. (2009). Lip-reading aids word recognition most in moderate noise: A Bayesian explanation using high-dimensional feature space. *PloS one*, *4*(3), 4638. doi:10.1371/journal.pone.0004638

MacDonald, J., Andersen, S., & Bachmann, T. (2000). Hearing by eye: Just how much spatial degradation can be tolerated? *Perception, 29*, 1155–1168. doi:10.1068/p3020

MacLeod, A., & Summerfield, Q. (1987). Quantifying the contribution of vision to speech perception in noise. *British Journal of Audiology*, *21*(2), 131-141. doi:10.3109/03005368709077786

Marques, L. M., Lapenta, O. M., Costa, T. L., & Boggio, P. S. (2016). Multisensory integration processes underlying speech perception as revealed by the McGurk illusion. *Language, Cognition and Neuroscience, 31*(9), 1115-1129. doi:10.1080/23273798.2016.1190023

Massaro, D. W. (1987). Categorical partition: A fuzzy-logical model of categorization behavior. In S. Harnad (Ed.), *Categorical perception: The groundwork of cognition* (pp. 254-283). New York, NY, US: Cambridge University Press.

Massaro, D. W., & Cohen, M. M. (1983). Evaluation and integration of visual and auditory information in speech perception. *Journal of Experimental Psychology: Human Perception and Performance, 9*(5), 753. doi:10.1037/0096-1523.9.5.753

Massaro, D. W., & Oden, G. C. (1980). Speech perception: A framework for research and theory. In N. J. Lass (Ed.), *Speech and language* (pp. 129-165). New York: Elsevier.

Massaro, D. W., & Palmer Jr., S. E. (1998). *Perceiving talking faces: From speech perception to a behavioral principle*. Cambridge, Massachusetts: MIT Press.

Massaro, D. W., and Cohen, M. M. (1993) Perceiving asynchronous bimodal speech in consonant-vowel and vowel syllables. *Speech Communication, 13*, 127–34. doi:10.1016/0167-6393(93)90064-R

Massaro, D. W., Cohen, M. M., Gesi, A., Heredia, R., & Tsuzaki, M. (1993). Bimodal speech perception: An examination across languages. *Journal of Phonetics, 21*(4), 445–478. Retrieved from https://psycnet.apa.org

Mastrantuono, E., Saldaña, D., & Rodríguez-Ortiz, I. R. (2017). An Eye Tracking Study on the Perception and Comprehension of Unimodal and Bimodal Linguistic Inputs by Deaf Adolescents. *Frontiers in Psychology, 8*, 1044. doi:10.3389/fpsyg.2017.01044

McGettigan, C., Faulkner, A., Altarelli, I., Obleser, J., Baverstock, H., & Scott, S. K. (2012). Speech comprehension aided by multiple modalities: Behavioural and neural interactions. *Neuropsychologia*, *50*(5), 762-776. doi:10.1016/j.neuropsychologia.2012.01.010

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature, 264*(5588), 746-748. doi:10.1038/264746a0

Meredith M. A., & Stein, B. E. (1986). Spatial factors determine the activity of multisensory neurons in cat superior colliculus. *Cognitive Brain Research, 369*, 350–354. doi:10.1016/0006-8993(86)91648-3

Miller, J. (1988). A warning about median reaction time. *Journal of Experimental Psychology: Human Perception and Performance*, *14*(3), 539–543. https://doi.org/10.1037/0096-1523.14.3.539

Mitterer, H., & Reinisch, E. (2017). Visual speech influences speech perception immediately but not automatically. *Attention, Perception, and Psychophysics*, *79*(2), 660-678. doi:10.3758/s13414-016-1249-6

Moon, I. J., & Hong, S. H. (2014). What is temporal fine structure and why is it important?. *Korean Journal of Audiology*, *18*(1), 1. doi: 10.7874/kja.2014.18.1.1

Moore, B. C. (2003). Coding of sounds in the auditory system and its relevance to signal processing and coding in cochlear implants. *Otology and Neurotology*, *24*(2), 243-254. doi: 10.1097/00129492-200303000-00019

Morís Fernández, L., Macaluso, E., & Soto-Faraco, S. (2017). Audiovisual integration as conflict resolution: The conflict of the McGurk illusion. *Human Brain Mapping*, *38*(11), 5691-5705. doi:10.1002/hbm.23758

Munhall, K. G., Kroos, C., Jozan, G., & Vatikiotis-Bateson, E. (2004). Spatial frequency requirements for audiovisual speech perception. *Perception and Psychophysics*, *66*(4), 574-583. doi:10.3758/BF03194902

Nath, A. R., & Beauchamp, M. S. (2012). A neural basis for interindividual differences in the McGurk effect, a multisensory speech illusion. *Neuroimage, 59*(1), 781-787. doi:10.1016/j.neuroimage.2011.07.024

National Institute of deafness and other communication disorders (2017) *Statistics about hearing*. Retrieved from https://www.nidcd.nih.gov/health/statistics/quick-statistics-hearing

Neely, K. K. (1956). Effect of visual factors on the intelligibility of speech. *The Journal of the Acoustical Society of America, 28*(6), 1275-1277. doi:10.1121/1.1908620

Newman, J. L., & Cox, S. J. (2012). Language identification using visual features. *IEEE Transactions on audio, speech, and language processing*, *20*(7), 1936-1947. doi: 10.1109/TASL.2012.2191956

Owens. E., & Blazek, B. (1985). Visemes observed by hearing impaired and normal-hearing adult viewers. *Journal of Speech, Language, and Hearing Research*, *28*, 381-393. doi:10.1044/jshr.2803.381

Oxenham, A. J. (2008). Pitch perception and auditory stream segregation: Implications for hearing loss and cochlear implants. *Trends in Amplification*, *12*(4), 316-331. doi:10.1177/1084713808325881

Paré, M., Richler, R. C., ten Hove, M., & Munhall, K. G. (2003). Gaze behavior in audiovisual speech perception: The influence of ocular fixations on the McGurk effect. *Perception and Psychophysics, 65*(4), 553-567. doi:10.3758/BF03194582

Peckens, C. A., & Lynch, J. P. (2013). Utilizing the cochlea as a bio-inspired compressive sensing technique. *Smart Materials and Structures*, *22*(10), 105027. doi:10.1088/0964-1726/22/10/105027

Peele, J. E. & Sommers, M. S. (2015). Prediction and constraint in audiovisual speech perception. *Cortex, 68*, 169-181. doi:10.1016/j.cortex.2015.03.006

Peelle J. E. (2018). Listening Effort: How the Cognitive Consequences of Acoustic Challenge Are Reflected in Brain and Behavior. *Ear and Hearing*, *39*(2), 204–214. doi:10.1097/AUD.0000000000000494

Peelle, J. E., & Sommers, M. S. (2015). Prediction and constraint in audiovisual speech perception. *Cortex, 68*, 169-181. https://doi.org/10.1016/j.cortex.2015.03.006

Peiffer, A. M., Mozolic, J. L., Hugenschmidt, C. E., and Laurienti, P. J. (2007). Age-related multisensory enhancement in a simple audiovisual detection task. *Neuroreport* 18, 1077–1081. doi: 10.1097/WNR.0b013e3281e72ae7

Pilling, M., & Thomas, S. (2011). Audiovisual cues and perceptual learning of spectrally distorted speech. *Language and speech*, *54*(4), 487-497. doi:10.1177/0023830911404958

Pimperton, H., Ralph-Lewis, A., & MacSweeney, M. (2017). Speechreading in deaf adults with cochlear implants: Evidence for perceptual compensation. *Frontiers in Psychology*, *8,* 106. doi:10.3389/fpsyg.2017.00106

Politzer-Ahles, S., & Pan, L. (2019). Skilled musicians are indeed subject to the McGurk effect. *Royal Society Open Science, 6*(4), 181868. doi:10.1098/rsos.181868

Proverbio, A. M., Massetti, G., Rizzi, E., & Zani, A. (2016). Skilled musicians are not subject to the McGurk effect. *Scientific reports, 6*, 30423. doi:10.1038/srep30423

Proverbio, A. M., Raso, G., & Zani, A. (2018). Electrophysiological indexes of incongruent audiovisual phonemic processing: unraveling the McGurk effect. *Neuroscience, 385*, 215-226. doi:10.1016/j.neuroscience.2018.06.021

Putzar, L., Hötting, K., & Röder, B. (2010). Early visual deprivation affects the development of face recognition and of audio-visual speech perception. *Restorative Neurology and Neuroscience, 28*(2), 251-257. doi: 10.3233/RNN-2010-0526

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. doi:10.1121/1.1579009

Raine, C. (2013). Cochlear implants in the United Kingdom: Awareness and utilization. *Cochlear Implants International, 14*(1), 32-37. doi: 10.1179/1467010013Z.00000000077

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience, 2*(1), 79. doi:10.1038/4580

Ratcliff, R. (1993). Methods for dealing with reaction time outliers. *Psychological Bulletin*, 114, 510-532.

Recanzone, G. H., & Sutter, M. L. (2008). The biological basis of audition. *Annual Review of Psychology*, *59*, 119-142. doi: 10.1146/annurev.psych.59.103006.093544

Reisberg, D., McLean, J., & Goldfield, A. (1987). Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. In B. Dodd & R. Campbell (Eds.), *Hearing by eye: The psychology of lip-reading* (pp. 97-113). Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Remez, R. E., Rubin, P. E., Pisoni, D. B., & Carrell, T. D. (1981). Speech perception without traditional speech cues. *Science, 212*(4497), 947-949. doi: 10.1126/science.7233191

Richard, F. D., Bond Jr., C. F., & Stokes-Zoota, J. J. (2003). One Hundred Years of Social Psychology Quantitatively Described. *Review of General Psychology, 7*(4), 331–363. doi:10.1037/1089-2680.7.4.331

Richardson, C. K., Bowers, D., Bauer, R. M., Heilman, K. M., & Leonard, C. M. (2000). Digitizing the moving face during dynamic displays of emotion. *Neuropsychologia*, *38*(7), 1028-1039. doi:10.1016/S0028-3932(99)00151-7

Robinson, C. W., Chandra, M., & Sinnett, S. (2016). Existence of competing modality dominances. *Attention, Perception, and Psychophysics, 78*(4), 1104-1114. doi:10.3758/s13414-016-1061-3

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. doi: 10.1121/1.428215

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. doi:10.1121/1.428215

Rosenblum, L. D. (2019). Audiovisual Speech Perception and the McGurk Effect. In *Oxford Research Encyclopedia of Linguistics*.

Rosenblum, L. D., Johnson, J. A., & Saldaña, H. M. (1996). Point-light facial displays enhance comprehension of speech in noise. *Journal of Speech, Language, and Hearing Research, 39*(6), 1159-1170. doi:10.1044/jshr.3906.1159

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

Rouger, J., Fraysse, B., Deguine, O., & Barone, P. (2008). McGurk effects in cochlear-implanted deaf subjects. *Brain Research, 1188*, 87-99. doi:10.1016/j.brainres.2007.10.049

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, 104*(17), 7295-7300. doi:10.1073/pnas.0609419104

Rubinstein, J. T. (2004). How cochlear implants encode speech. *Current Opinion in Otolaryngology and Head and Neck Surgery, 12*(5), 444-448. doi: 10.1097/01.moo.0000134452.24819.c0

Saalasti, S., Tiippana, K., Kätsyri, J., & Sams, M. (2011). The effect of visual spatial attention on audiovisual speech perception in adults with Asperger syndrome. *Experimental Brain Research, 213*(2-3), 283-290. doi:10.1007/s00221-011-2751-7

Saldaña, H. M., & Rosenblum, L. D. (1993). Visual influences on auditory pluck and bow judgments. *Perception and Psychophysics, 54*(3), 406-416. doi:10.3758/BF03205276

Sekiyama, K., and Tokhura, Y. (1991). McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. *Journal of the Acoustical Society of America, 90*, 1797–1825. doi:10.1121/1.401660

Sekiyama, K., and Tokhura, Y. (1993). Inter-language differences in the influence of visual cues in speech perception. *Journal of Phonetics, 21*, 427–444. Retrieved from https://psycnet.apa.org

Sekiyama, K., Soshi, T., & Sakamoto, S. (2014). Enhanced audiovisual integration with aging in speech perception: A heightened McGurk effect in older adults. *Frontiers in Psychology*, *5*, 323. doi:10.3389/fpsyg.2014.00323

Sekuler, R., Sekuler, A. B., & Lau, R. (1997). Sound changes perception of visual motion. *Nature, 384*, 308-309. Retrieved from https://www.nature.com

Senju, A., & Johnson, M. H. (2009). The eye contact effect: mechanisms and development. *Trends in Cognitive Sciences*, *13*(3), 127-134. doi:10.1016/j.tics.2008.11.009

Shams, L., & Kim, R. (2010). Crossmodal influences on visual perception. *Physics of Life Reviews*, *7*(3), 269-284. doi:10.1016/j.plrev.2010.04.006

Shams, L., Kamitani, Y., & Shimojo, S. (2000). Illusions: What you see is what you hear. *Nature, 408*(6814), 788. doi:10.1038/35048669

Shams, L., Kamitani, Y., & Shimojo, S. (2001). Sound modulates visual evoked potentials in humans. *Journal of Vision, 1*(3), 479-479. doi:10.1167/1.3.479

Shannon, R. V., Zeng, F. G., Kamath, V., Wygonski, J., & Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science, 270*(5234), 303-304. doi:10.1126/science.270.5234.303

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.

Shimojo, S., & Shams, L. (2001). Sensory modalities are not separate modalities: Plasticity and interactions. *Current Opinion in Neurobiology, 11*(4), 505-509. doi:10.1016/S0959-4388(00)00241-5

Skinner, M. W., Ketten, D. R., Holden, L. K., Harding, G. W., Smith, P. G., Gates, G. A., Neely, G. J., Kletzker, G. R., Brunsden, B., & Blocker, B. (2002). CT-derived estimation of cochlear morphology and electrode array position in relation to word recognition in Nucleus-22 recipients. *Journal of the Association for Research in Otolaryngology*, *3*(3), 332-350. Doi:10.1007/s101620020013

Skipper, J. I., Van Wassenhove, V., Nusbaum, H. C., & Small, S. L. (2007). Hearing lips and seeing voices: how cortical areas supporting speech production mediate audiovisual speech perception. *Cerebral Cortex, 17*(10), 2387-2399. doi:10.1093/cercor/bhl147

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. doi:10.1121/1.3158835

Stacey, P. C. (2006). *Studies of auditory training to improve speech perception by adult cochlear-implant users* (Doctoral dissertation, University of York, York). Retrieved from https://www.york.ac.uk/library/collections/theses/

Stacey, J. E., Howard, C. J., Mitra, S. & Stacey P. C. (under review). Audio-visual integration in noise: Influence of auditory and visual stimulus degradation on eye-movements and perception of the McGurk effect.

Stacey, P.C., & Summerfield, A.Q. (2007). Effectiveness of computer-based auditory training in improving the perception of noise-vocoded speech. *Journal of the Acoustical Society of America, 121*, 2923-2935. doi:10.1121/1.2713668

Stacey, P.C., Kitterick, P.T., Morris, S.D. and 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. doi:10.1016/j.heares.2016.04.002

Stacey, P.C., Raine, C.H, O'Donoguhe, G.M., Tapper, L., Twomey, T., & Summerfield, A.Q. (2010). Effectiveness of computer-based auditory training for adult users of cochlear implants. *International Journal of Audiology, 49*, 347-356. doi:10.3109/14992020903397838

Stickney, G. S., Nie, K., & Zeng, F. G. (2005). Contribution of frequency modulation to speech recognition in noise. *The Journal of the Acoustical Society of America*, *118*(4), 2412-2420. doi:10.1121/1.2031967

Strand, J., Cooperman, A., Rowe, J., & Simenstad, A. (2014). Individual differences in susceptibility to the McGurk effect: Links with lipreading and detecting audiovisual incongruity. *Journal of Speech, Language, and Hearing Research, 57*(6), 2322-2331. doi:10.1044/2014_JSLHR-H-14-0059

Stropahl, M., Schellhardt, S., & Debener, S. (2017). McGurk stimuli for the investigation of multisensory integration in cochlear-implant users: The Oldenburg Audio Visual Speech Stimuli (OLAVS). *Psychonomic Bulletin & Review, 24*(3), 863-872. doi:10.3758/s13423-016-1148-9

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. doi:10.1121/1.1907309

Summerfield, Q. (1992). Lipreading and audio-visual speech perception. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, *335*(1273), 71-78. doi:10.1098/rstb.1992.0009

Thomas, S. M., & Jordan, T. R. (2002). Determining the influence of Gaussian blurring on inversion effects with talking faces. *Perception and Psychophysics, 64*, 932–944. doi:10.3758/BF03196797

Tiippana, K. (2014). What is the McGurk effect?. *Frontiers in Psychology*, *5*, 725. doi:10.3389/fpsyg.2014.00725

Tremblay, C., Champoux, F., Voss, P., Bacon, B. A., Lepore, F., & Théoret, H. (2007). Speech and non-speech audio-visual illusions: A developmental study. *PloS one*, *2*(8), 742. doi:10.1371/journal.pone.0000742

Tuomainen, J., Andersen, T. S., Tiippana, K., & Sams, M. (2005). Audio–visual speech perception is special. *Cognition*, *96*(1), B13-B22. doi:10.1016/j.cognition.2004.10.004

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

Tye-Murray, N., Sommers, M. S., & Spehar, B. (2007b). The effects of age and gender on lipreading abilities. *Journal of the American Academy of Audiology*, *18*(10), 883-892. doi:10.3766/jaaa.18.10.7

Tye-Murray, N., Sommers, M. S., Spehar, B., Myerson, J., & Hale, S. (2010). Aging, audiovisual integration, and the principle of inverse effectiveness. *Ear and Hearing*, *31*(5), 636–644. doi:10.1097/AUD.0b013e3181ddf7ff

Tye-Murray, N., Spehar, B., Myerson, J., Hale, S., & Sommers, M. S. (2016). Lipreading and audiovisual speech recognition across the adult lifespan: Implications for audiovisual integration. *Psychology and Aging*, *31*(4), 380–389. doi:10.1037/pag0000094

Tye-Murray, N., Spehar, B., Sommers, M., & Barcroft, J. (2016). Auditory training with frequent communication partners. *Journal of Speech, Language, and Hearing Research*, *59*(4), 871-875. doi:10.1044/2016_JSLHR-H-15-0171

Ujiie, Y., Asai, T., & Wakabayashi, A. (2015). The relationship between level of autistic traits and local bias in the context of the McGurk effect. *Frontiers in Psychology*, *6*, 891. doi:10.3389/fpsyg.2015.00891

Ujiie, Y., Asai, T., Tanaka, A., Asakawa, K., & Wakabayashi, A. (2014). Autistic traits predict weaker visual influence in the McGurk effect. *Personality and Individual Differences*, *60*, 51-52. doi:10.1016/j.paid.2013.07.211

UK Cochlear Implant Group (2004). Criteria of Candidacy for Unilateral Cochlear Implantation in Postlingually Deafened Adults I: Theory and Measures of Effectiveness. *Ear and Hearing, 25* (4), 310–335. doi: 10.1097/01.AUD.0000134549.48718.53

Van Engen, K. J., Xie, Z., & Chandrasekaran, B. (2017). Audiovisual sentence recognition not predicted by susceptibility to the McGurk effect. *Attention, Perception, and Psychophysics, 79*(2), 396–403. doi:10.3758/s13414-016-1238-9

Van Engen, K., Dey, A., Sommers, M., & Peelle, J. E. (In press). Audiovisual speech perception: Moving beyond McGurk.

van Wassenhove, V. (2013). Speech through ears and eyes: Interfacing the senses with the supramodal brain. *Frontiers in Psychology, 4,* 388. doi:10.3389/fpsyg.2013.00388

van Wassenhove, V., Grant, K. W. & Poeppel (2005). Visual speech speeds up the neural processing of auditory speech. *Proceedings of the National Academy of Sciences of the United States of America, 102,* 1181-1186. doi:10.1073/pnas.0408949102

van Wassenhove, V., Grant, K. W. and Poeppel, D. (2007). Temporal window of integration in bimodal speech, Neuropsychologia 45, 598–607.

Vatikiotis-Bateson, E., Eigsti, I. M., Yano, S., & Munhall, K. G. (1998). Eye movement of perceivers during audiovisual speech perception. *Perception and Psychophysics, 60*(6), 926-940. doi:10.3758/BF03211929

Walden, B. E., Erdman, S. A., Montgomery, A. A., Schwartz, D. M., & Prosek, R. A. (1981). Some effects of training on speech recognition by hearing-impaired adults. *Journal of Speech, Language, and Hearing Research, 24*(2), 207-216. doi:10.1044/jshr.2402.207

Walker, J. T., & Scott, K. J. (1981). Auditory–visual conflicts in the perceived duration of lights, tones, and gaps. *Journal of Experimental Psychology: Human Perception and Performance, 7*(6), 1327. doi:10.1037/0096-1523.7.6.1327

Welch, R. B. & Warren, D. H. (1986) Intersensory interactions. In K. R. Boff, L. Kaufman & J. P. Thomas, (Eds.), *Handbook of Perception and Human Performance* (pp.25.1-25.36). New York: Wiley.

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. doi: 10.1121/1.2773993

Wifall, T., Hazeltine, E., & Mordkoff, J. T. (2016). The roles of stimulus and response uncertainty in forced-choice performance: an amendment to Hick/Hyman Law. *Psychological Research, 80*(4), 555-565. doi:10.1007/s00426-015-0675-8

Wilson, A. H., Alsius, A., Paré, M., & Munhall, K. G. (2016). Spatial frequency requirements and gaze strategy in visual-only and audiovisual speech perception. *Journal of Speech, Language, and Hearing Research*, *59*(4), 601-615. doi:10.1044/2016_JSLHR-S-15-0092

Witten, I. B., & Knudsen, E. I. (2005). Why seeing is believing: merging auditory and visual worlds. *Neuron, 48*(3), 489-496. doi:10.1016/j.neuron.2005.10.020

Wouters, J., McDermott, H. J., & Francart, T. (2015). Sound coding in cochlear implants: From electric pulses to hearing. *IEEE Signal Processing Magazine*, *32*(2), 67-80. doi: 10.1109/MSP.2014.2371671

Wouters, J., McDermott, H. J., & Francart, T. (2015). Sound coding in cochlear implants: From electric pulses to hearing. *IEEE Signal Processing Magazine, 32*(2), 67-80. doi: 10.1109/MSP.2014.2371671

Wozniak, V. D., & Jackson, P. L. (1979). Visual vowel and diphthong perception from two horizontal viewing angles. *Journal of Speech, Language, and Hearing Research, 22*(2), 354-365. doi:10.1044/jshr.2202.354

Xu, L. (2016). Temporal Envelopes in Sine-Wave Speech Recognition. In W. Hess, & M. Cooke (Eds.). *Interspeech* (pp. 1682-1686). International Speech Communication Association. doi:10.21437/Interspeech.2016-171

Yi, A., Wong, W., & Eizenman, M. (2013). Gaze patterns and audiovisual speech enhancement. *Journal of Speech, Language, and Hearing Research*. doi:10.1044/1092-4388(2012/10-0288)

Zekveld, A. A., Kramer, S. E., & Festen, J. M. (2011). Cognitive load during speech perception in noise: The influence of age, hearing loss, and cognition on the pupil response. *Ear and Hearing, 32*(4), 498-510. doi: 10.1097/AUD.0b013e31820512bb

Appendix A

Experiment 1 stimuli

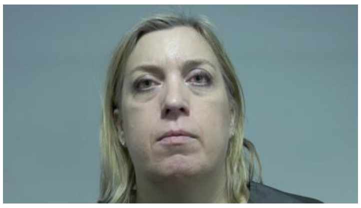
Each talker had three congruent stimuli (/ba:/, ga:/ & da:/) and two tokens of the incongruent stimuli $A_{BA}V_{GA.}$ The two tokens were two different instances of the talker uttering the syllable.

Talker

1

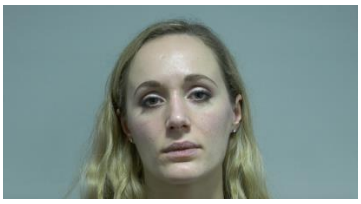
2

3



4

5



6

7

8

Appendix B

Pilot study to determine levels of noise

The purpose of the pilot was to ascertain which levels of background noise were appropriate to use in the eye-tracking experiments. Congruent stimuli (BA, GA, DA) were presented from 4 talkers in auditory and visual blurring: clear speech in white noise; vocoded speech in white noise and visual blur. Figure B1 shows participants' performance in each of these conditions (percentage correct) represented by the black circles. For each condition the data point which was in the middle of high degradation and no degradation was chosen as the moderate level of noise, the 2 levels of noise used for experiments 2 and 3 are indicated in by the grey squares. Chance level was at 33%.

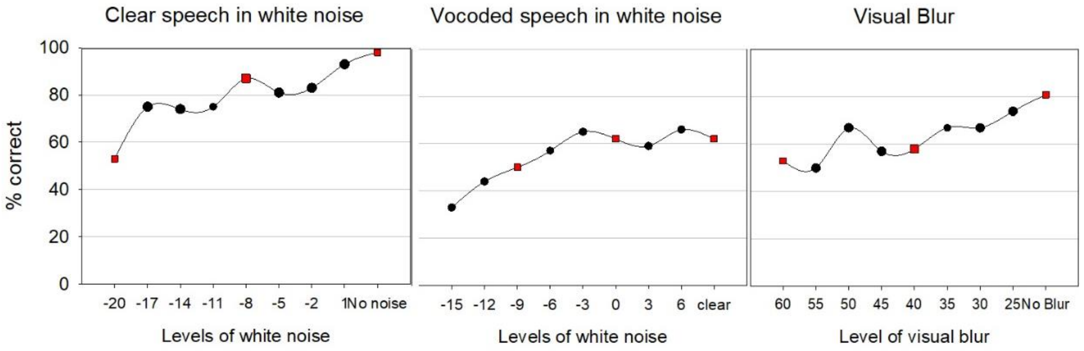


Figure B1 Accuracy in different levels and types of noise