The impact of visual speech on neural processing of auditory speech

Aisling O'Sullivan, BA., MA.

Under the supervision of

Dr. Edmund C. Lalor

Prof. Richard R. Reilly

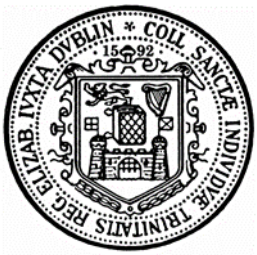
A dissertation submitted to the

University of Dublin, Trinity College

In fulfilment of the requirements for the degree of

Doctor of Philosophy

January 2021



Department of Electronic and Electrical Engineering

University of Dublin, Trinity College

Declaration

I, Aisling O'Sullivan, confirm that this thesis has not been submitted as an exercise for a degree at this or any other university and is entirely my own work.

I agree to deposit this thesis in the University's open access institutional repository or allow the Library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

I consent to the examiner retaining a copy of the thesis beyond the examining period, should they so wish (EU GDPR May 2018).

Signed,

_____

Aisling O'Sullivan

January 2021

Summary

Communication is central to our society and enables us to build relationships with others. Conversation is, most of the time, quite effortless, yet the neural activity that underpins this function has puzzled scientists for decades. Speech is a very complex and fast changing signal, and since no two speakers talk in the same way, it is a challenging feat for the brain to rapidly interpret and comprehend what a speaker is saying.

There is a large body of research, dating back to over 50 years ago, that has explored the processing of speech sounds from the vibration of the ear drum, all the way to the comprehension of the meaning of that sound, which recruits extensive regions of cortex. This work has led to a general consensus that speech is processed in a hierarchical manner, whereby simple acoustic properties such as amplitude and frequency information are processed in lower stages of the hierarchy and at higher stages there is specificity for complex features of speech sounds. This hierarchy is thought to anatomically spread from the anterior (lower stages) to posterior (higher stages) region of the temporal lobe bilaterally.

It is also well known that vision impacts auditory speech perception, with a growing body of evidence pointing to extensive interactions between the visual and auditory systems, from eye movements impacting ear drum movement (Gruters *et al.*, 2018), up to cortical regions that are particularly sensitive to both mouth movements and speech sounds (Zhu & Beauchamp, 2017). Nevertheless, there remains many questions about the interactions between auditory and visual speech signals at each stage of speech processing, and how these interactions may flexibly adapt to changes in environment such as the level of background noise and the reliability of the visual information.

Tackling these questions has been challenging. One reason for this, is that the vast majority of experiments to date have been constrained to repeated presentations of isolated syllables and words, which limit interpretations of how a system built to deal with the rapid dynamics of continuous speech really works. This has led to a drive towards the use of natural speech in experiments (Hamilton & Huth, 2018). In this thesis, we use natural speech along with recently developed analysis techniques (Crosse *et al.*, 2016a; de Cheveigne *et al.*, 2018) in order to examine the influence of visual speech on the cortical tracking of continuous, dynamic auditory speech (Luo *et al.*, 2010; Golumbic *et al.*, 2013b; Crosse *et al.*, 2015a; Crosse *et al.*, 2016b).

In the first study, we examine how the relevance of the visual speech affects the neural

tracking of the auditory speech. We use multivariate regression to show that a model trained on the neural activity when the visual speech is relevant differs from a model trained on neural responses to speech when the visual speech is irrelevant and distracting. We also find that parieto-occipital alpha power is substantially increased when the visual speech is irrelevant. This suggests that there is a suppression of visual speech processing when the visual speech is irrelevant compared with when it matches the attended auditory speech.

In the second study, our goal was to isolate multisensory interactions by comparing the combined unisensory (A+V) responses with the responses to the multisensory (AV) speech. We used a spectrogram and phonetic feature representation of the speech, in order to test for multisensory interactions at different stages of speech processing. We found multisensory interactions at both stages of speech processing, which is in line with current models proposing a multistage integration framework (Peelle & Sommers, 2015). We then added noise to the acoustic signal in order to test if interactions between the auditory and visual speech would adapt to the degraded listening conditions. In this case, we again found multisensory effects at both stages of speech processing, however, the integration effects at the phonetic stage of processing were greater than when there was no noise added to the acoustics. This suggests a possible greater reliance on visual articulatory information when the acoustic speech is degraded.

We then designed an experiment to remove the visual articulatory details by blurring the mouth of the speaker and assessed its effect on perception and our neural indices of multisensory integration. Using the envelope, spectrogram and phonetic feature representations of the speech, we found that blurring the mouth had the greatest influence on phonetic encoding, with a significant drop in performance of the phonetic model when the mouth was blurred. However, we found no differences in multisensory interactions across the two conditions for the envelope model, the spectrogram model or the phonetic feature model.

Together, the work in this thesis contributes to our understanding of the cortical system that processes natural audiovisual speech. These studies provide a framework for studying EEG responses to continuous speech and allowed us to address several questions related to the influence of visual speech on the processing of auditory speech. Future studies which build directly off the results reported in this thesis are also proposed in detail and we hope that this work will help future studies to explore multisensory integration effects using natural speech.

Acknowledgements

I would like to thank the many people whose support and advice throughout this work was invaluable to me.

Firstly, Ed, for being a patient and supportive supervisor and for always being welcoming to talk to. Thanks for your kindness and generosity, and for the many opportunities to travel. I have really enjoyed my time in the lab.

Richard for all your help and advice through the years, particularly all your help with the Irish lab.

Adam for teaching me so much throughout the whole PhD, for all the memorable adventures and for keeping my work-life balance in tow. Also thanks for reading my work and always providing constructive criticism.

Andy for your listening ear, insightfulness and wisdom, Kevin for all the climbing, cycling and pure craic, Aaron for all the chats and friendship, Lauren for bringing a blast of colour in more ways than one! The workday always felt shorter when all of you were around and never as fun with one of you missing! Tiziana for your statements, opinions and for always making me laugh, and Federico for simply being yourself… (yes, yourself!).

I would also like to thank all the past and present members of the Lalor lab: Mick, Giovanni, Denis, Michael, Nate, Emily, Shy and Fahrin.

Most importantly, thanks to my dad and all my family for their invaluable support through the years.

This research was kindly supported by Science Foundation Ireland.

Table of Contents

Publications Arising from this Thesis

Journal Articles

- **O'Sullivan, A. E.**, Lim, C. Y., & Lalor, E. C. (2019). Look at me when I'm talking to you: Selective attention at a multisensory cocktail party can be decoded using stimulus reconstruction and alpha power modulations. *European Journal of Neuroscience*, 00: 1– 14. https://doi.org/10.1111/ejn.14425

Journal Articles in Review

- **O'Sullivan, A.**, Crosse, M. J., Di Liberto, G. M., de Cheveigné, A., Lalor, E. C. (2020). Neurophysiological indices of audiovisual speech integration are enhanced at the phonetic level for speech in noise.

Journal Articles in Preparation

- **O'Sullivan, A.**, Lalor, E. C. The contribution of visual dynamic and articulatory cues to audiovisual speech perception in noise.
- Nidiffer, A. R., Cao, Z., Syzmula, L., **O'Sullivan, A.**, Lalor, E. C. Enhanced visual phonetic encoding of unheard speech in visual cortex for trained lipreaders
- Ahmed F., **O'Sullivan, A.**, Nidiffer, A. R., Zuk, N. J., Lalor, E. C. The integration of audio and visual speech in a cocktail party environment depends on attention,

Conference Poster Presentations

- **O'Sullivan, A., E**, Nidiffer A. R., & Lalor, E. C. Assessing the contribution of visual speech features to audiovisual speech perception in noise. *Society for Neuroscience 2019*, Chicago, USA

- **O'Sullivan, A., E**, Nidiffer A. R., & Lalor, E. C. Assessing the contribution of visual speech features to audiovisual speech perception in noise. *Advances and Perspectives in Auditory Neuroscience 2019*, Chicago, USA

- **O'Sullivan, A., E**, & Lalor, E. C. Multisensory effects along the speech processing hierarchy in quiet and noisy environments. *Society for Neuroscience 2018*, San Diego, USA

- **O'Sullivan, A., E**, & Lalor, E. C. Multisensory effects along the speech processing hierarchy in quiet and noisy environments. *Advances and Perspectives in Auditory Neuroscience 2018*, San Diego, USA

- **O'Sullivan, A., E**, Crosse M., J., Di Liberto G., M., Majeski J., de Cheveigné A., & Lalor, E. C. Indexing Multisensory Integration of Natural Speech using Canonical Correlation Analysis *International Multisensory Research Forum 2018, Toronto, Canada.*

- **O'Sullivan, A., E**, & Lalor, E. C. Decoding Attentional Selection at a Multisensory Cocktail Party. *Society for Neuroscience 2017*, Washington DC, USA

- **O'Sullivan, A., E**, & Lalor, E. C. Decoding Attentional Selection at a Multisensory Cocktail Party. *Advances and Perspectives in Auditory Neuroscience 2017*, Washington DC, USA

- **O'Sullivan, A., E**, Crosse M., J. & Lalor, E. C. Assessing multisensory effects in natural audiovisual speech processing using EEG predictions. *PIRE workshop on Hierarchical Multisensory Integration 2017, Barcelona, Spain*

- **O'Sullivan, A., E**, Crosse M., J. & Lalor, E. C. Assessing multisensory effects in natural audiovisual speech processing using EEG predictions. *Auditory Cortex 2017, Banff, Canada.*

- **O'Sullivan, A., E**, Crosse M., J. & Lalor, E. C. Assessing multisensory effects in natural audiovisual speech processing using EEG predictions. *PIRE workshop on Hierarchical Multisensory Integration 2017, Barcelona, Spain*

- **O'Sullivan, A., E**, Crosse M., J. & Lalor, E. C. Assessing multisensory effects in natural audiovisual speech processing using EEG predictions. *International Multisensory Research Forum 2017, Nashville, USA.*

Glossary of Acronyms

AC	Auditory Cortex
AV	Audiovisual
ECoG	Electrocoticography
EEG	Electroencephalography
ERP	Event-Related Potential
fMRI	functional Magnetic Resonance
HRTF	Head-Related Transfer Function
IC	Inferior Colliculus
IFC	Inferior Frontal Cortex
LTI	Linear Time Invariant
MEG	Magnetoencephalography
MSI	Multisensory integration
MMN	Mismatch Negativity
pSTG	Posterior Superior Temporal Gyrus
ROI	Region of Interest
STS	Superior Temporal Sulcus
TRF	Temporal Response Function
TVSA	Temporal Visual Speech Area
VOT	Voice Onset Time

Chapter 1.

Introduction

The ability to comprehend language is central to human interaction. For the majority of people, hearing alone enables language comprehension. The most common use of language however, is through face-to-face conversations, and seeing the speaker undoubtedly contributes to our understanding of what they are saying. For those with imperfect hearing, being able to see the speaker substantially improves comprehension over hearing alone, and for others, particularly those that are deaf, lipreading and sign language are the means through which language comprehension is achieved.

This identifies two important systems in the brain involved in language comprehension and communication – the auditory and visual systems. For those with normal hearing ability, the population on which this thesis is based, seeing the face of the speaker becomes more beneficial when the listening conditions are challenging, e.g., due to background noise or people talking nearby.

How this benefit comes about however, remains incompletely understood. While research on speech perception dates back to over 60 years ago, it was initially focused on speech as an auditory percept of an acoustic stimulus (French & Steinberg, 1947; Klatt, 1979; Samuel, 2011). One major challenge in this field is to understand how the variable acoustic input is mapped to discrete speech units such as phonemes, syllables and words. Insights from speech perception studies have shed light on this question, and have generated a prominent model which suggests that speech is processed in a hierarchical manner with increasing encoding specificity and invariance from posterior to anterior temporal areas (Hickok & Poeppel, 2007; DeWitt & Rauschecker, 2012). However, many questions remain, and in particular the role and impact of visual speech on this hierarchical system remains an open question (Campbell, 2008; Peelle & Sommers, 2015; Vatikiotis-Bateson & Munhall, 2015; Holler & Levinson, 2019).

One way to address this question is to identify what information is contained in the visual speech. It has been shown that the visual speech signal itself contains information that is correlated with the amplitude of the acoustic waveform (i.e., the facial movements that produce the speech sound have a temporal relationship with that speech sound), and also complementary information whereby the shape of the mouth and

position of the tongue and the lips can help disambiguate between acoustically similar speech sounds which are susceptible to masking in noisy environments. The challenge then remains to identify how this information from visual speech, interacts with the brain's speech processing pipeline, such that it influences the identity of speech units.

A recent perspective on this question proposed that the two abovementioned forms of visual information – correlated and complementary – interact with two different stages of auditory processing. It was proposed that the correlated information impacts early processing of the speech sound and that the complementary information impacts a later stage of processing in constraining syllable and word identification (Peelle & Sommers, 2015). Direct evidence for this multistage integration framework however is lacking for a number of reasons. Firstly, the vast majority of research to date on this topic has used isolated words or syllables that lack the dynamics of natural speech – making it difficult to disentangle contributions from correlated and complementary components of visual speech. Secondly, only a few experiments have sought to isolate specific hierarchical speech representations in neurophysiological data, and none that we are aware of have examined how those different representations might be affected by visual input. Finally, not many studies have tested how integration at different stages may vary as a function of the listening conditions and the available visual information. In order to address these questions, we build on recent approaches that model electroencephalography (EEG) responses to natural speech using both acoustic and phonetic feature representations (Di Liberto *et al.*, 2015) as well as recently developed approaches for indexing multisensory integration in natural audiovisual speech experiments (Crosse *et al.*, 2015a; Crosse *et al.*, 2016b). The combination of hierarchical speech representations with analysis techniques that allow the use of continuous, natural speech provides a suitable framework in which the abovementioned questions can be tackled in a way that contributes to our understanding of audiovisual speech perception as we experience it in every-day life. This is the goal of this thesis, with the following specific aims set out below.

1.1. Aims

The overarching aim of this work is to investigate the neural signatures of the impact of visual speech on auditory processing of natural speech. To do this, we set out 3 aims which align with the studies described in chapters 3-5 of this thesis.

1. Examine how cortical tracking of auditory speech is affected by attention to visual speech in a multisensory cocktail party

2. Investigate the effect of listening conditions on the interaction of auditory and visual speech at different stages of processing
3. Examine how the availability of different features of the visual speech (i.e., mouth motion versus mouth details) impacts the integration of audiovisual speech

In summary, we examine how cortical tracking of auditory speech changes when the accompanying visual speech is relevant or irrelevant. We also investigate how visual speech impacts different stages of auditory processing by isolating neural indices of multisensory integration of natural audiovisual speech. This is done for speech in quiet and noisy speech in order to assess how the impact of visual speech may vary with the quality of the acoustic information. Finally, we examine how modifying the availability of specific visual speech information impacts our measures of multisensory integration. This is done with the goal of better understanding how the different aspects of visual speech (e.g., mouth motion versus mouth details) contribute to audiovisual speech processing.

1.2. Thesis outline

Chapter 2 describes the neurophysiology of auditory speech, visual speech and mechanisms of audiovisual speech integration. This is followed by a description of the imaging method used in this work, electroencephalography (EEG), and discusses the analysis frameworks that are used to process the EEG data.

In Chapter 3, a study on the influence of visual speech on the cortical tracking of the acoustic envelope as a function of the relevance of the visual speech is presented and discussed. The main finding of this study is that attention to visual speech modulates cortical activations when attending to speech in a multisensory cocktail party environment.

Chapter 4 presents a novel approach for indexing multisensory integration effects using natural speech. It also investigates the impact of visual speech at 2 different stages of auditory speech processing and tests how these measures are affected when listening conditions are degraded. It is found that at both spectrogram and phonetic stages of processing there are significant multisensory effects in quiet and noisy listening conditions. Furthermore, when the speech is degraded the multisensory effect is enhanced at the level of phonetic processing relative to speech in quiet. This result suggests a greater reliance on visual articulatory features when the speech is noisy.

Chapter 5 explores the differential contribution of mouth dynamics versus mouth details to the multisensory integration of audiovisual speech. Specifically, it examines whether these visual speech features contribute differently to the 2 stages of auditory speech processing explored in chapter 4. We find that phonetic encoding is worse when the mouth is blurred compared with when it is clear, whereas the spectrogram encoding is similar in both conditions. This result provides neural evidence for the mouth details providing visual phonetic cues such as place and manner of articulation.

Finally, chapter 6 presents a discussion of the main findings of the research carried out in this thesis and outlines future work that builds upon the foundations of this work.

Chapter 2.

Background

This chapter provides a background to the work in this thesis, and is divided into four sections. The first section describes current theories on the neurophysiology of speech perception. The following section focuses on the role of visual information in speech perception, while the third section provides background on how the brain combines information from the auditory and visual modalities, with a particular focus on audiovisual speech. The fourth and final section then presents an overview of the neuroimaging and analysis techniques used to study neural processing of speech and multisensory integration.

2.1. The neurophysiology of auditory speech perception

One prominent theory of speech perception is that auditory speech is processed in a hierarchical manner, whereby each stage is specialised for processing particular speech features. In this section we first describe the subcortical processing of sound and then discuss the role of cortex with a focus on speech sounds.

2.1.1. Auditory periphery and subcortical processing of sound

The peripheral auditory system transforms sound pressure waves into neural firings through the auditory nerve. When sound waves reach the ear, they travel through the ear canal and cause the tympanic membrane (i.e., ear drum) to vibrate. The vibrations of the tympanic membrane are then transferred to the oval window through three small interconnected middle ear bones, or ossicles (i.e., the malleus, incus, and stapes). The sound then travels to the cochlea of the inner ear, within which the pressure waves are transformed into neural signals. The cochlea itself consists of a small coiled cavity with tonotopic organisation (frequency-to-place mapping), which performs frequency analysis of the incoming sound wave by splitting it up into multiple frequency-bands (Yang *et al.*, 1992). This tonotopic organisation is conserved throughout much of the auditory system, including the auditory cortex, suggesting that it is important for processing natural sounds.

The ability of the cochlea to transform the pressure wave into neural signals is due to tiny mechanosensory cells known as hair cells. The hair cells have protruding organelles called stereocilia which respond to the shear force generated by the sound

pressure wave as it moves through the cochlea. The hair cell movement results in a change in voltage across the hair cells which are connected to the spiral ganglion neurons of the auditory nerve.

After the incoming sound waves are transduced into neural activity by the peripheral auditory system, they travel to the cochlear nucleus (CN) via the auditory nerve. The CN innervates the superior olive complex and the lateral lamiscus. The neural signals then target the inferior colliculus (IC) which is a major convergence region in the auditory pathway receiving feed-forward inputs from the abovementioned subcortical structures, as well as feed-back connections from auditory cortex (Winer *et al.*, 1998; Stebbings *et al.*, 2014 for a review)

Within the IC, three functionally distinct regions have been defined (Oliver & Morest, 1984): the central nucleus, which is predominantly auditory (Aitkin *et al.*, 1994); the lateral (or external) cortex, which is multisensory (Aitkin *et al.*, 1978), and receives considerable non-auditory input (Oliver & Morest, 1984); and the dorsal cortex, which receives a large proportion of descending (top-down) projections from the auditory cortex (for detailed review on the IC see: Schreiner & Winer, 2005).

The neural signals then propagate to the medial geniculate body (MGB), which is also thought to be involved in integration of auditory features as well as multisensory integration (Schreiner & Winer, 2005). MGB activity has also been shown to be modulated by task (von Kriegstein *et al.*, 2008), and the context, i.e., stimulus specific adaptation (Antunes *et al.*, 2010). After MGB, the auditory neural pathway then reaches the auditory cortex. For a more detailed discussion on the auditory subcortical structures and pathways see Schreiner and Winer (2005) and Purves *et al.* (2008).

Although it is well known that cortex plays a central role in speech processing, there is evidence to suggest that preliminary processing of the sound carried out by subcortical structures contributes to the cortical processing of speech (Pannese *et al.*, 2015). In particular, the cochlear nucleus has been shown to be sensitive to the voice onset time (VOT) of syllables in rats (Clarey *et al.*, 2004) and so may play a similar role of coding VOT in humans. The inferior colliculus is thought to be particularly sensitive to spectrotemporal variations that are inherent in communication signals, and is also thought to play a critical role in auditory learning (Bajo *et al.*, 2010; Chandrasekaran *et al.*, 2012). Furthermore, the finding of responses that phase-lock to the fundamental frequency of vocalizations in MGB (Wallace *et al.*, 2007) along with its sensitivity to particular combinations of spectral features (Kanwal & Rauschecker, 2007) has also led to

suggestions of the importance of MGB in analysis of auditory communication signals. Nevertheless, cortex plays a substantial role in speech processing and the topic of this thesis deals only with cortical processing of speech. A discussion on the role of cortex in speech processing is presented in the next section.

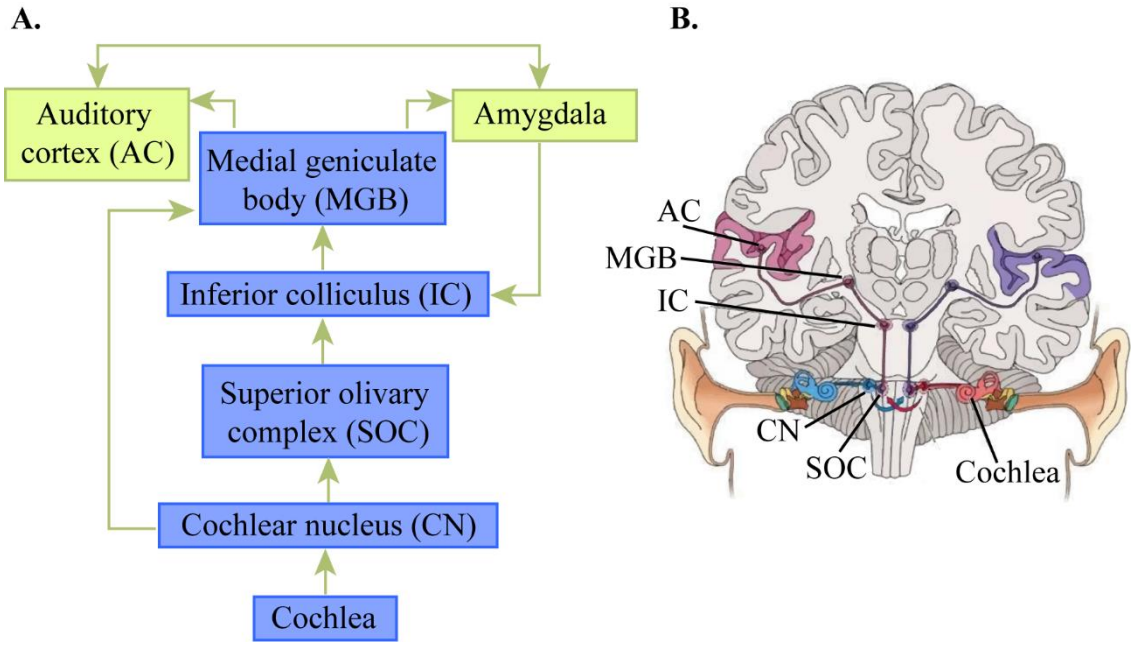


Figure 2.1. The subcortical auditory pathway. A. Schematic of the subcortical model for auditory communication. The major ascending projections are shown, and absence of arrows between regions does not imply absence of anatomical and/or functional connection, adapted from (Pannese et al., 2015). B. Diagram of the auditory pathways shows the location of the structures shown in A, adapted from (Purves et al., 2008).

2.1.2. Cortical processing of speech

After travelling from the ear and up through the various subcortical structures, a neural representation of the sound arrives at the auditory cortex. The core region of auditory cortex (located in Heschyl's gyrus) is thought to have a simple receptive-field organisation, and is involved in processing spectro-temporal fluctuations in the sound wave (Howard III *et al.*, 1996; Nourski *et al.*, 2014; Nourski, 2017). Then, 'higher' areas (i.e., the superior temporal sulcus and gyrus, STS and STG) are thought to combine or integrate information received from the lower cortical areas, resulting in these 'higher' areas having increasing selectivity for more complex sound properties (Fig. 2.2; Binder *et al.*, 2000; Hickok & Poeppel, 2007; Rauschecker & Scott, 2009; de Heer *et al.*, 2017).

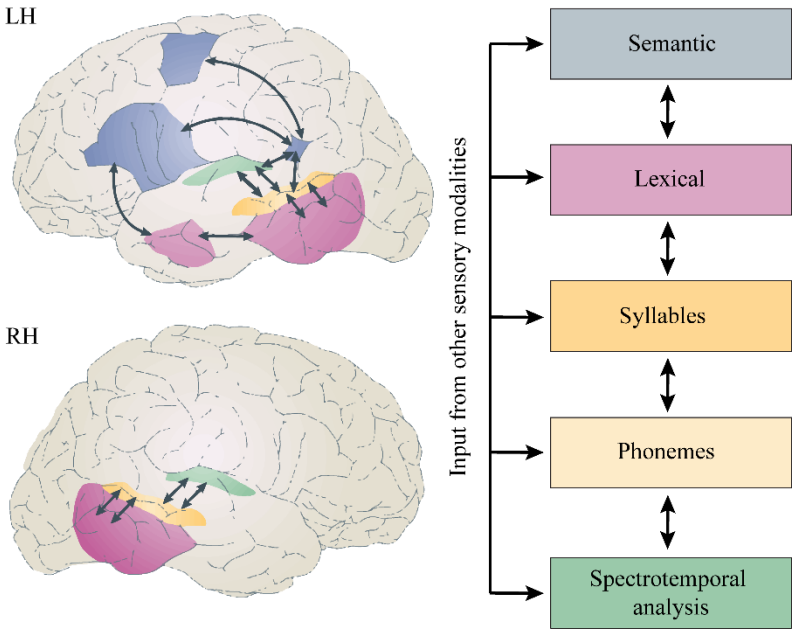


Figure. 2.2. Functional anatomical model of auditory speech processing hierarchy. This depicts the main functional blocks of the hierarchical model of language (right) and their corresponding cortical location in which they are processed (left). Adapted from Hickok and Poeppel (2007).

In the context of speech processing, it is interesting that the neural responses in these higher areas beyond primary auditory cortex have been shown to be more invariant to differences in spectro-temporal content, resulting in a grouping or categorisation of spectro-temporal features. These spectro-temporal categories actually closely correspond to different features of the produced speech (Liebenthal *et al.*, 2005; Chang *et al.*, 2010; Liebenthal *et al.*, 2010; Chan *et al.*, 2013; Mesgarani *et al.*, 2014; Hamilton *et al.*, 2018; for review see: Yi *et al.*, 2019), leading to the suggestion that these higher level regions are particularly important for mapping spectro-temporal information onto speech units.

The smallest speech unit is called a phoneme and is defined by linguists as the smallest linguistic unit that changes the meaning of a word in a particular language (e.g., /d/ and /m/ as in date vs. mate; Jakobson *et al.*, 1951; Chomsky & Halle, 1968). Phonemes can be mapped onto a set of phonological features in order to describe how speech sounds are articulated by humans (see Fig. 2.3; Chomsky & Halle, 1968). These features include properties of consonants, such as place of articulation (i.e., the location in the mouth where the constriction and obstruction of air occurs), manner of articulation (i.e., configuration and interaction of speech organs, such as tongue, lips, and palate), and glottal state of sounds (e.g., vibration, aspiration), as well as properties of vowels (e.g., back, high, and rounded vowels). Perception of these speech units from the speech sound waveform is thought to be a critical step in understanding speech (Liberman *et al.*, 1967; Mesgarani *et al.*, 2008; Overath *et al.*, 2015). This task is particularly difficult though,

due to the fact that the spectral content of an individual phoneme can vary depending on differences in speaker anatomy and physiology (Fant, 1966; Klatt & Klatt, 1990), differences in the listening conditions, such as noise or reverberation (Houtgast & Steeneken, 1973) or depending on the phoneme's location in a word (Öhman, 1966; Kent & Minifie, 1977) and varies with changes in speaking rate (Gay, 1978; Miller & Baer, 1983). Thus, a simple spectrotemporal profile does not define any one phoneme since this profile can vary with the abovementioned factors, highlighting the challenge of speech recognition.

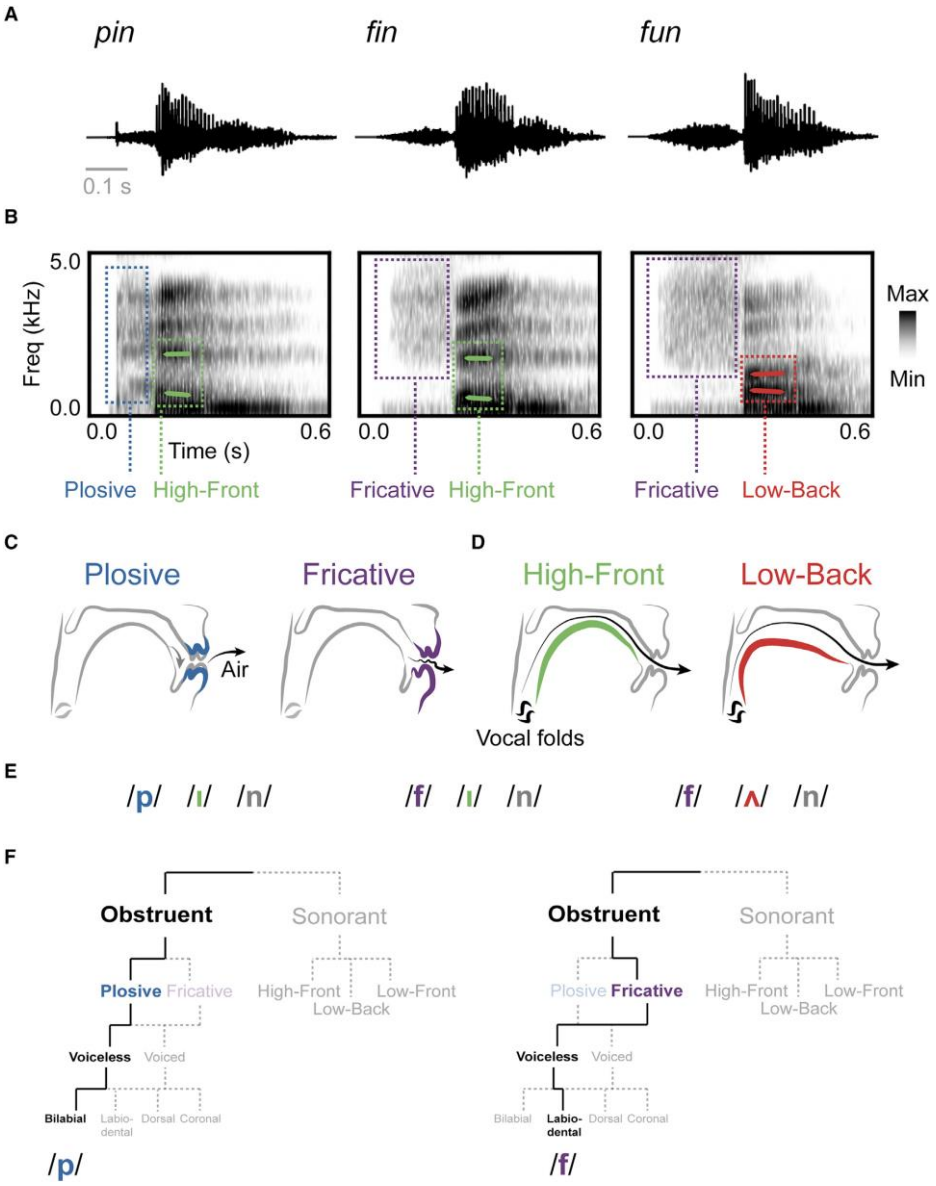


Figure 2.3. Speech sounds can be described in a number of different ways. A. The acoustic waveform, B. The spectro-temporal profile of each word. C-D. The different positions of the vocal apparatus in producing the example speech sounds. F. Linguistic method of describing the features of the produced sound. Adapted from Yi et al. (2019).

Evidence of responses which display categorisation of phonological features in STG (Formisano *et al.*, 2008; Chang *et al.*, 2010; Lee *et al.*, 2012; Mesgarani *et al.*, 2014;

see Fig. 2.4) suggests a way in which speech-specific analysis may take place along the auditory pathway. In order to comprehend speech, these spectro-temporal features then need to be integrated to form lexical (i.e., word) and semantic (i.e., meaning) representations. STG has also been implicated in representations at the lexical level from findings that acoustic-phonetic responses in STG are strongly influenced by context – in particular, the responses are affected by learned language statistics such as phoneme probability and word frequency (Cibelli *et al.*, 2015; Leonard *et al.*, 2015; Leonard *et al.*, 2016). Since the response in STG is modified by the preceding input, it is likely that STG does not just represent a response to an instantaneous input, but instead maintains information for a certain period of time. In line with this, it was recently shown that responses to phonemes are maintained longer when the lexical uncertainty is higher (Gwilliams *et al.*, 2020), suggesting that phonetic information is maintained until lexical ambiguity is resolved. STG and the surrounding superior temporal area is likely to play a major role in this. Finally, access to meaning of the perceived word has been shown to involve extensive regions of cortex, suggesting that the semantic representation of words and sentences is distributed across a large network rather than being restricted to a particular brain region (Mitchell *et al.*, 2008; Huth *et al.*, 2016; Anderson *et al.*, 2017; Pereira *et al.*, 2018).

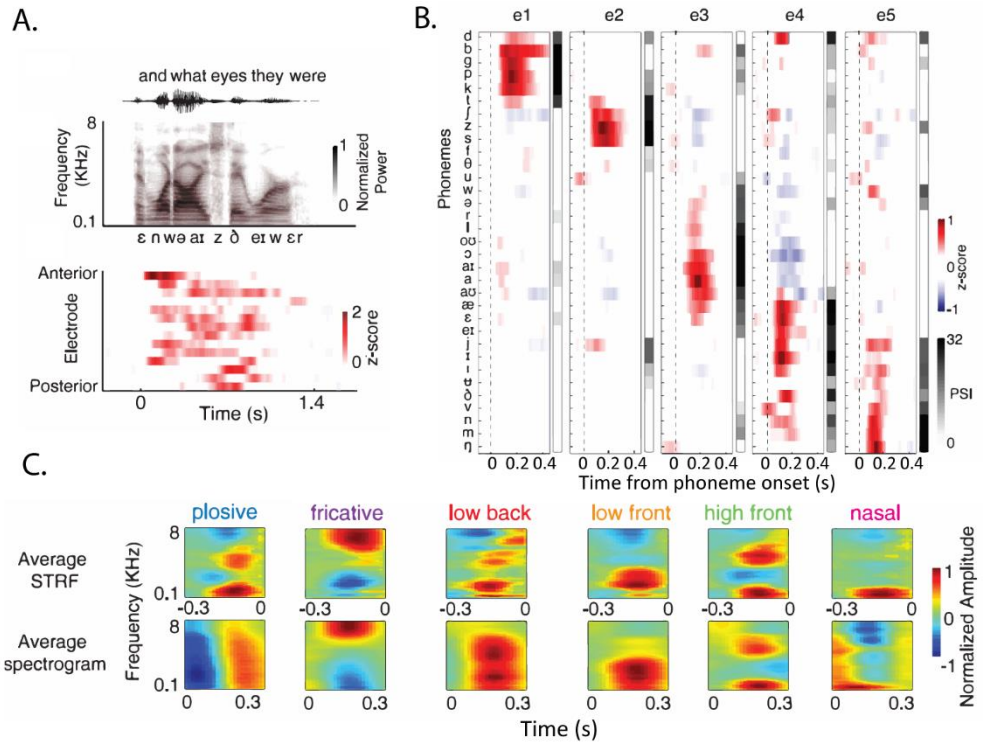


Figure 2.4. STG responses to speech. A. Example exert from stimulus, showing the acoustic waveform, the spectrogram and the phoneme transcription for that sentence. The neural responses to that sentence are also shown with electrodes on the y-axis and time on the x-axis. B. Average electrode responses for 5 electrodes show that certain

electrodes are particularly sensitive for certain phonemes. C. Sensitivity of STG to particular spectro-temporal patterns that correspond to phonetic features. Adapted from Mesgarani et al. (2014).

Together, the research to date suggests that the cortical processing of speech follows a hierarchical structure, where each 'section' along this hierarchy has a computationally distinct function. It is unlikely however, that any of these processes follow a strictly feedforward manner, from smaller units (e.g., phonemes) to larger units (e.g., syllables and words) given that the human cortex is characterized by a massive amount of anatomical feedback versus feedforward connections (Rasmussen, 1964). Feedback connections are suspected to play a role in modulating excitability of neurons in lower regions of the hierarchy based on predictions of the expected input (Kuperberg & Jaeger, 2016).

Our review so far has focused on how the brain encodes and represents auditory speech. The next section will deal with the contribution of visual speech to this typically thought of auditory percept.

2.2. The role of visual speech in audiovisual speech perception

Although we can usually understand speech quite well from the acoustics alone, it is well known that the visual information from the speaker's face significantly impacts our perception of the acoustics (Sumby & Pollack, 1954; Reisberg *et al.*, 1987). One common example of this is the ventriloquist effect, which creates an illusion that the voice appears to be coming from the mouth of the moving puppet, or when we go to the cinema and perceive the voice to be coming from the actor's mouth even though the sound is coming from the speakers on either side of the screen. Another example, which was stumbled upon by two scientists performing an experiment for a different purpose, is the McGurk effect (McGurk & Macdonald, 1976). This is an effect of perceiving an illusory phoneme when an incongruent auditory and visual phoneme are presented simultaneously. Altogether, these examples point out the profound effect of visual inputs on auditory speech perception. The following sections present possible mechanisms through which these perceptual influences may occur.

2.2.1. The correlated visual speech information

The ability of visual speech to impact auditory speech processing is thought to be driven by a spatial and/or temporal correspondence between the auditory and visual signals. One well-known temporal relationship is the correspondence between the face movements of

the producer and the speech sound waveform which is produced. Specifically, it has been shown that the mouth area is correlated with the acoustic envelope of the speech over a broad range of frequencies (Fig. 2.5). Other features of the face such as the eyebrows, jaw and chin movements have also been shown to be correlated with the acoustics (Jiang *et al.*, 2002; Yehia *et al.*, 2002). Head movements alone have also been shown to improve speech perception, possibly through their alignment with the speaker's voice and may convey information about stress and prominence, i.e., prosody (Munhall *et al.*, 2004). These correlations can benefit speech encoding particularly when the speech is difficult to understand, for e.g., due to noise in the acoustics. The stage of auditory processing at which the visual lip movements contributes remains unknown. This question will be discussed further in sections 2.2.4 and 2.3.

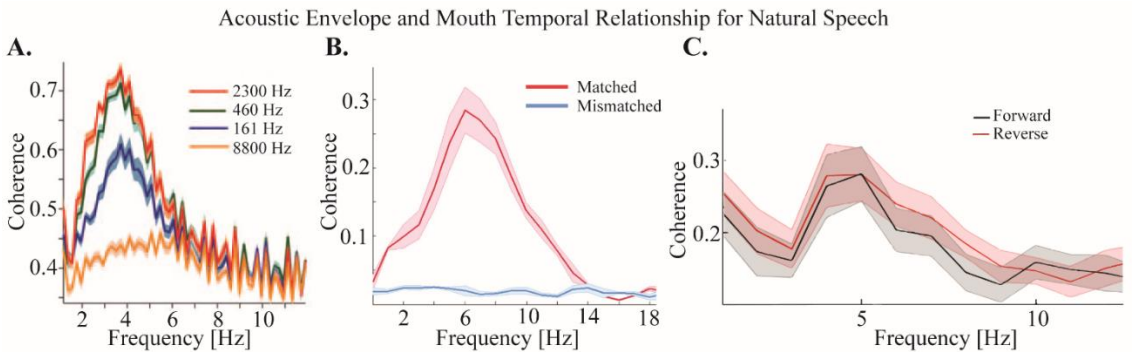


Figure 2.5. The temporal relationship between mouth movements and the acoustic waveform of speech shows consistency across a variety of natural speech stimuli. A. The coherence between difference frequency bands of the acoustic waveform and mouth area, adapted from Chandrasekaran et al. (2009). B. The coherence between the speech envelope and the mouth area for matching versus mismatching audiovisual speech signals, adapted from Park et al. (2016). C. The coherence between the envelope and the mouth area for forward and time reversed speech has a similar profile, adapted from Hauswald et al. (2018).

2.2.2. The complementary visual speech information

Besides visual speech being correlated with auditory speech, which provides a redundant form of information, it also provides complementary information about the acoustics. This is because any particular mouth movement made by the talker is compatible with only a few auditory phonemes (Neti *et al.*, 2000; Cappelletta & Harte, 2012, Fig. 2.6).

This can be readily seen for example with the speech sounds /p/ and /d/, which are acoustically different in F2 formant space which is easily masked in noisy environments (Miller and Nicely, 1955) and with impaired hearing (Walden et al., 1975). However, visually there is a lip-puff for /p/ which is absent for /d/ and this can allow us to use visual speech to distinguish between the words "map" and "mad" when the listening conditions

are degraded.

To characterise the information in visual speech, visual speech features can be grouped into categories termed ‘visemes’ in line with their acoustic analogue phonemes (Woodward & Barber, 1960; Fisher, 1968). There is not a one-to-one mapping between phonemes and visemes, and instead it is the case that many phonemes can be grouped as having similar visual speech features. An example of this is the bilabials /p/ and /b/ which are visually identical but acoustically different since /b/ is voiced and /p/ un-voiced (Fig. 2.6).

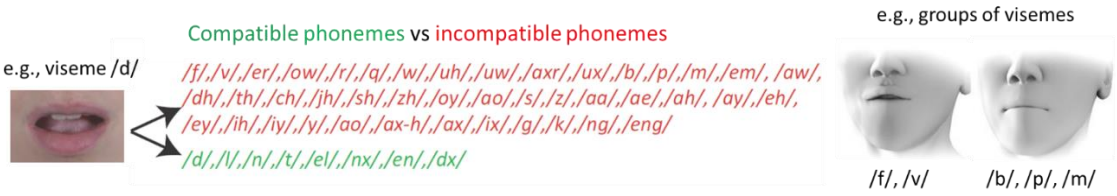


Figure 2.6. The relationship between acoustic speech features and visual speech features. Left: For a given visual speech feature there are only a small number of compatible acoustic speech features, adapted from Karas et al. (2019). Right: An example of two different groupings of visually similar phonemes, i.e., visemes. The bilabials (/b/, /p/, /m/), and the fricatives (/f/, /v/).

The advantage of having auditory and visual streams of information present, results in a reduction in the uncertainty of the number of possible phoneme candidates that reside in *both* visual and audio groups. This was demonstrated by the findings that neither the number of auditory only nor the visual only phonetic candidates related to the audiovisual performance, instead it was determined by the number of candidates in their intersection, where a higher intersection density resulted in poorer recognition (Tye-Murray *et al.*, 2007). This suggests that visual and auditory phonetic information is combined in order to benefit audiovisual speech recognition.

2.2.3. The temporal relationship between auditory and visual speech

Another property of visual speech that can benefit auditory speech perception is that it sometimes leads the auditory speech by some time (approx. ranges from 60-300 ms, see: Chandrasekaran *et al.*, 2009; Schwartz & Savariaux, 2014) and it is thought that this property of visual speech allows information about the speech content to be extracted before the acoustics begin. This was recently examined in an experiment that presented subjects with audiovisual words, where one set of AV words were mouth-leading and the other set were voice-leading. They found that only in the case of mouth-leading words did the neural response in pSTG differ from the response to the word in the auditory-only presentation. Specifically, they found that the response was supressed for mouth-leading

words over voice-leading words. The authors' explanation for this was that the leading visual speech information enhances neural responses to phonemes compatible with that visual speech and suppresses responses to phonemes which would be incompatible with that visual speech (see Fig. 2.7). Since there are more incompatible than compatible phonemes then the overall effect is a suppression of the response (Karas *et al.*, 2019). This suggests that the benefit provided by visual-leading audiovisual speech may be to constrain the number of possibilities of the upcoming phoneme.

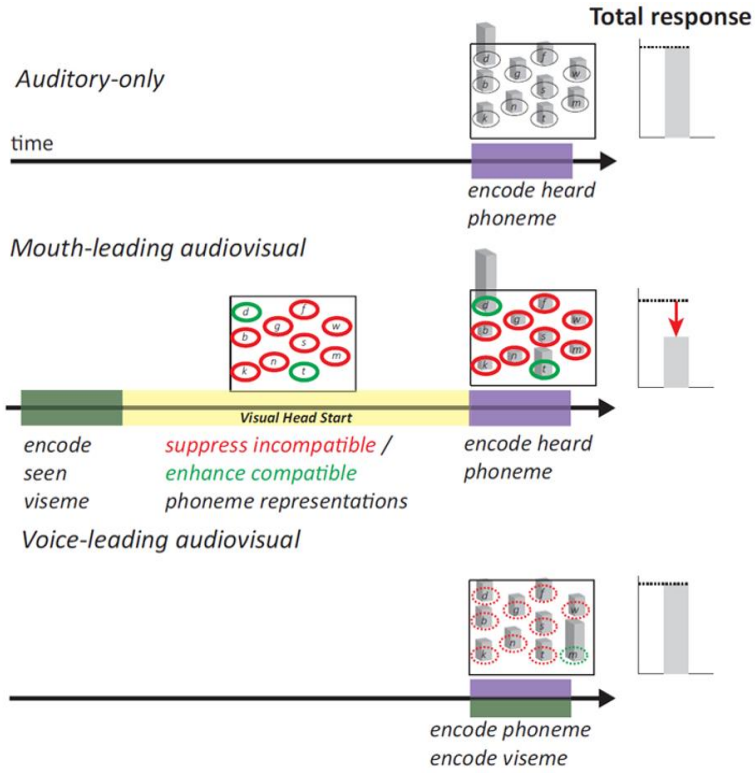


Figure 2.7. A diagram of the impact of visual-leading audiovisual speech on phoneme encoding. Top: For the auditory only case neurons which encode similar acoustic features to the sound presented respond. Middle: For mouth-leading audiovisual speech there is a suppression of responses to incompatible phonemes and an enhancement of responses to compatible phonemes which results in an overall reduction in response – since there are less compatible than incompatible phonemes. Bottom: For the case of voice-leading speech there is no early suppression effect and so the response is similar to the response to auditory only speech. Adapted from (Karas et al., 2019).

Another way in which visual speech is thought to influence auditory speech processing is the so-called 'attention in time' hypothesis. The basis for this comes from the reasoning that since visual speech sometimes leads the auditory speech onset, this head start in time may alert auditory cortex that a sound is coming soon (Schroeder *et al.*, 2008; Tye-Murray *et al.*, 2011; ten Oever *et al.*, 2014). However, Strand and colleagues (2019) showed that a visual stimulus which provided information about the onset, amplitude and offset of the acoustic speech did not improve speech recognition for words

or sentences, and Schwartz *et al.* (2004) found that modulating a rectangle by lip movements did not improve speech intelligibility. Thus, it seems that the 'alerting' of auditory cortex by a visually leading stimulus does not account for the improved intelligibility observed for visual leading AV speech vs auditory leading AV speech (Karas *et al.*, 2019).

Instead, it appears that it is required that the visual stimulus provides some information about the upcoming acoustics (although it is not always required to match the acoustics, e.g., the McGurk effect) in order to benefit speech intelligibility. One recent example of this was where researchers showed that the area of a disc was sufficient to provide frequency information about the upcoming acoustics, whereby narrow and long discs were associated with higher frequency sounds and broad and round discs were matched with lower frequency sounds (Plass et al., 2019). Importantly, when these disc shapes were mismatched they did not provide any perceptual benefit. In a similar manner, biological motion presented using point-light displays has also been shown to benefit auditory speech perception in noise (Rosenblum *et al.*, 1996). Altogether, this suggests that it is required for the visual stimulus to provide some information about the acoustic signal in order to benefit perception/recognition.

2.2.4. Neural pathways and mechanisms of visual speech processing

While there is a wealth of studies examining auditory and audiovisual speech perception, there is much less known about visual speech perception in normal hearing people (i.e., people not explicitly trained at lipreading). One key question is how visual speech itself is processed. One idea is that visual speech is processed in auditory cortex and auditory association cortex by mapping it to an auditory representation, another idea however, is that visual cortex processes speech as speech and has become specialised for speech processing analogous to auditory cortex. The evidence for each of these possibilities is discussed in detail below.

Neuroimaging literature on lipreading shows widespread activity in both ventral and dorsal visual pathways in response to visual speech (for review see: Bernstein & Liebenthal, 2014), but it remains unknown what these activations represent. There is no doubt that the physical visual signal that is received through the eyes is mapped from the retina to primary visual cortex. Somewhere further up the processing pathway however, the visual input is recognised as coming from the same object as the auditory signal that has entered through the ears. The precise role of the visual system in this process remains

unclear. Specifically it is unknown to what extent the visual system encodes the visual speech information, i.e. from light intensities, to motion, to the level of whole word recognition.

One idea is that our visual system processes low-level visual information such as light intensity and motion and then sends that information directly to auditory regions where it is processed as speech. This hypothesis implies that specialization for speech will not be found upstream in visual cortical areas. This idea has primarily been driven by studies on lipreading that have found auditory cortical activation (Sams *et al.*, 1991; Calvert *et al.*, 1997; Ludman *et al.*, 2000; Pekkola *et al.*, 2005; Besle *et al.*, 2008). These results have led to suggestions that visual regions feed information to early auditory cortical regions where the stimulus information is processed as though it had been transduced by the peripheral auditory system. Other studies have also shown auditory cortical activation by a visual stimulus, however, the activation occurred regardless of the category of visual stimulus (Kayser *et al.*, 2008) and auditory activation was also found for a somatosensory stimulus (Lemus *et al.*, 2010). Therefore it remains to be seen if this auditory response simply serves a modulatory function or if it is actually a representation of perceptual detail of the visual speech.

Recent evidence has shed light on this idea of auditory regions 'synthesising' speech from the silent videos by showing that auditory cortex actually tracks the dynamics of visual speech at very low frequencies (~0.5 Hz, Bourguignon *et al.*, 2020). The authors state that this is a result of auditory cortex forming some course-grained representation of the corresponding auditory speech. Nonetheless, this rate corresponds closely to the rate of sentences and so appears too slow for capturing individual words or even visual phonetic information – which is known to be visually salient (Weinholtz & Dias, 2016). Thus it is likely that these intermediate representations are formed in visual and heteromodal regions (Bernstein *et al.*, 2011).

Certainly, it is plausible to suggest that visual cortex itself would be capable of speech specific processing. However, there are few results in the literature that speak directly to how the levels of speech that can be perceived visually, are neurally represented. Some recent work examining this question has found that tracking of the unheard acoustic envelope in visual cortex was greater for forward played silent speech than when the speech was reversed (Hauswald *et al.*, 2018) suggesting some higher level speech processing present in visual regions. It was also shown that one can find a marker

of visual speech feature tracking (termed 'visemes') in visual cortex, above and beyond visual motion tracking (O'Sullivan *et al.*, 2017). Neuroimaging work has also found an enhancement of lip processing regions in visual cortex during lipreading (Ozker *et al.*, 2018) suggesting that visual regions may extract some visual phonetic information during silent speech. Together these provide some promising evidence of higher level speech processing taking place in visual cortex.

The location at which visual phonetic information may be encoded in visual cortex has been mainly investigated by contrasting BOLD responses to speech vs non-speech facial movements. One such study found that while speech gestures lead to activations in FFA, LOC and V5/MT, these activations are actually stronger for non-speech gestures, suggesting that these areas are not specialised for visual speech processing. Instead, it was shown that posterior MTG and STS ROIs were selective for speech. The localized posterior temporal speech selective area was dubbed the temporal visual speech area (TVSA) – see fig. 2.8 (Bernstein *et al.*, 2011). A later study from Files and colleagues (2013) employed a vMMN paradigm to isolate sensitivities to differences in visual speech features. They found that the left posterior temporal cortex (putative TVSA) was sensitive to perceptually far deviants (they used "zha" or "fa") whereas it was not sensitive to perceptually near deviants ("zha" or "ta"). This suggests that it is the left posterior temporal cortex which appears sensitive to violations of visual speech category, and so plays some role in encoding information about visual speech features. This information may then be fed forward into language processing areas such as pSTS/G for integration with the auditory speech information.

The STS and STG are well known sites of multisensory integration and pSTS has been shown to be sensitive to vocal mouth movements (Calvert *et al.*, 1997; Bernstein *et al.*, 2002; Calvert & Campbell, 2003) making it ideally suited to interact with visual speech areas and mediate the effects of visual speech input on auditory phonemic perception. A discussion of the research on audiovisual speech integration follows in the next section.

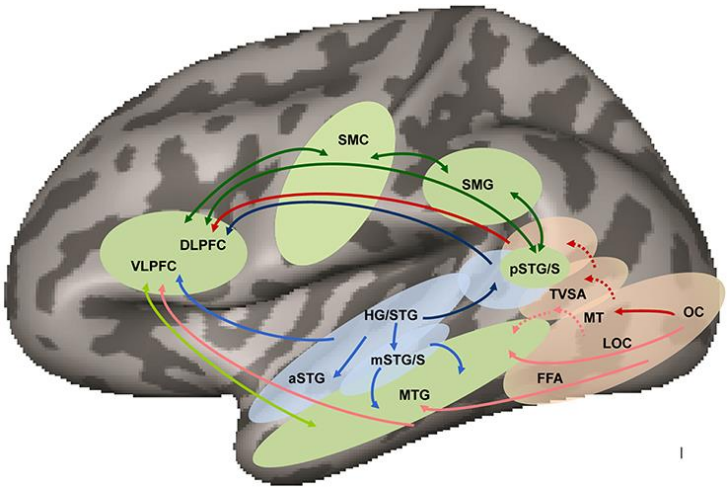


Figure 2.8. A functional anatomical model of audiovisual speech processing. The coloured arrows show the direction of information flow between regions. For example, the blue arrows show that auditory areas project to posterior temporal cortex, middle temporal gyrus and frontal regions. The red arrows show that visual regions project to both ventral (light red arrows) and dorsal (dark red arrows) regions of cortex. Between some areas there is bidirectional connectivity – shown by the bidirectional arrows. Adapted from (Bernstein & Liebenthal, 2014).

2.3. Multisensory integration

The fact that visual inputs impact auditory processing demonstrates that there is some interaction between these two sensory inputs. This section presents common approaches used to quantify multisensory integration effects, discusses possible underlying mechanisms of integration of audiovisual speech and lastly, talks about the automaticity of multisensory integration.

2.3.1. Quantifying multisensory effects and interactions in the brain

Measuring multisensory integration effects requires taking into account the processing of the two unimodal stimuli and isolating the independent processes involved that are due to multisensory interactions. Put another way, the unimodal processing first needs to be characterised and accounted for before obtaining estimates of multisensory processing. Many different models have been proposed for studying behavioural and neurophysiological effects of multisensory integration (for a review on these see: Stevenson *et al.*, 2014). Some of the earliest methods used for quantifying multisensory integration were based on recordings from single neurons in the SC of cats (Meredith & Stein, 1983), where multisensory integration was quantified in terms of the difference in spike counts when the stimulus was multisensory compared with the strongest unisensory response (Fig. 2.9, A-C), given by the equation:

$$MSI = \frac{AV - \max(A,V)}{\max(A,V)}$$

However, this approach is not suitable for non-invasive neuroimaging methods (such as EEG, MEG and fMRI) that record activity from large groups of neurons. This is because if neurons within that region respond to both unisensory stimuli then the maximum criterion would be greater than zero even if there were no multisensory responses in that region.

Another approach to deal with this issue is to instead test if the multisensory response is different from the sum of the two unisensory responses. This is referred to as the additive model (Barth *et al.*, 1995):

$$MSI = AV - (A + V)$$

In this case, the response is sub-additive when MSI < 0, additive when MSI = 0, and super-additive when MSI > 0 (Fig. 2.9, D-E). However, if there are common activations that are not directly related to sensory processing (e.g., motor activity) in both unisensory responses then that activity is effectively doubled in the summed unisensory response but is only present once in the multisensory response. This may explain why many multisensory responses exhibit response suppression (i.e., sub-additive) relative to the summed unisensory response (for a discussion on this see: Besle *et al.*, 2004b). To help alleviate this issue null trials can be used in which there is no sensory stimulus while subjects perform the same task as during real trials and the activity from these null trials is subtracted from all conditions in an attempt to remove the common (not stimulus related) activations.

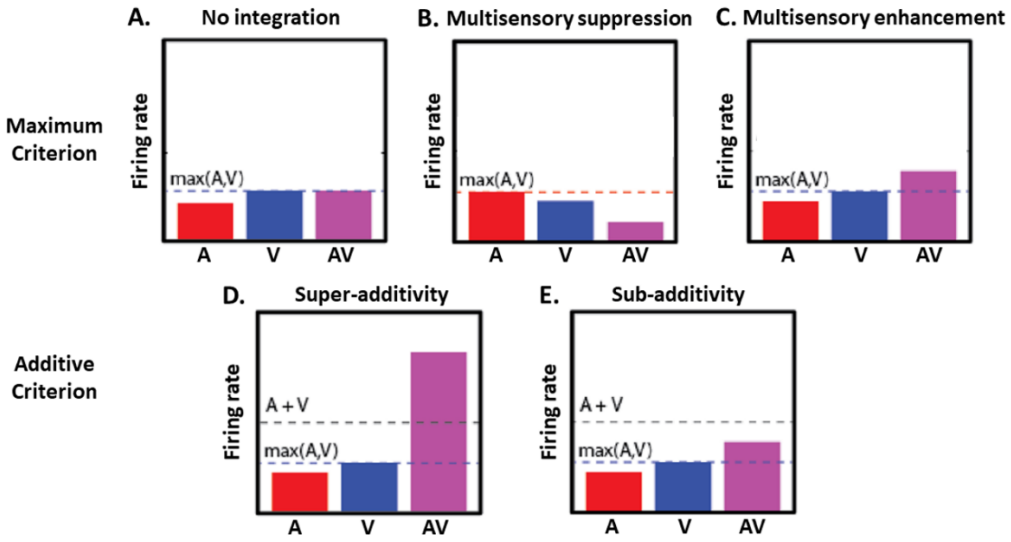


Figure 2.9. Illustration of maximum (top) and additive (bottom) criterion for quantifying multisensory integration. Adapted from (Stevenson et al., 2014)

2.3.2. Mechanisms of audiovisual speech integration

Current views on audiovisual speech integration suggest that it occurs at multiple stages of processing (van Wassenhove *et al.*, 2005; Arnal *et al.*, 2009; Peelle & Sommers, 2015), such that unimodal inputs are not processed separately first and then combined later, but instead, at the very early stages of processing there are cross modal influences on unisensory processing. Evidence for multisensory interactions at the earliest stages of processing have come from findings of auditory responses at the level of the brainstem being affected by visual speech input (Fig. 2.10; Musacchia *et al.*, 2006). Response enhancement in primary auditory cortex to audiovisual speech compared with auditory only speech has also been shown (Schroeder *et al.*, 2008; Okada *et al.*, 2013), as well as multisensory interactions in auditory cortex which precede those in STS (Möttönen *et al.*, 2004). However, it is thought that these early stage interactions are not speech-specific as similar effects have been reported for incongruent speech stimuli as well as non-speech stimuli (Klucharev *et al.*, 2003; Jääskeläinen *et al.*, 2004; Stekelenburg & Vroomen, 2007; Arnal *et al.*, 2009), and instead, these interactions appear to be dependent on visual motion preceding the auditory input. This influence of visual motion cues on the auditory response is thought to be facilitated by a direct corticocortical pathway from visual to auditory cortex (Arnal *et al.*, 2009).

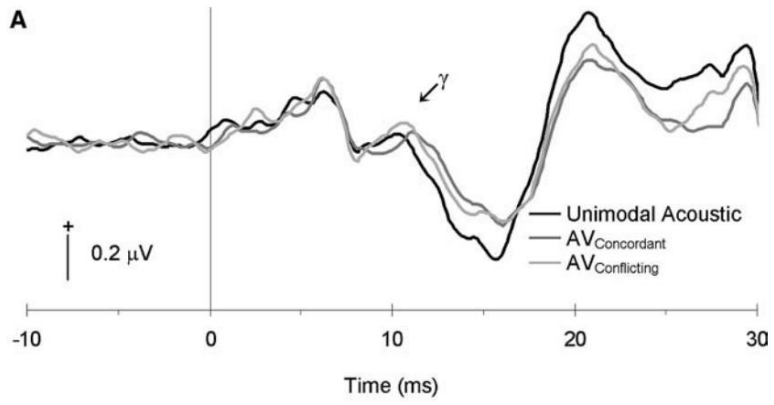


Figure. 2.10. Brainstem response to auditory and audiovisual speech syllables. The dark line shows the response to the auditory-only stimulus (unimodal acoustic). The dark grey line shows the responses to the audiovisual stimuli where they are congruent (concordant) and the lighter grey line is the response when they are incongurent (conflicting). Adapted from (Musacchia et al., 2006).

Regarding the later stage of integration, studies in non-human primates have found that STS receives inputs from both auditory and extrastriate visual cortex (Seltzer & Pandya, 1978), making it a suitable site for audiovisual integration. Unsurprisingly, there is a wealth of evidence demonstrating that STS is indeed involved in audiovisual

integration of both speech and non-speech stimuli (Calvert *et al.*, 2000; Sekiyama *et al.*, 2003; Wright *et al.*, 2003; Beauchamp *et al.*, 2004; Callan *et al.*, 2004; van Atteveldt *et al.*, 2004; Miller & D'Esposito, 2005; Stevenson & James, 2009; Werner & Noppeney, 2010). The precise role of STS in audiovisual speech integration is a topic of much interest. STS has shown to be sensitive to AV congruency, suggesting that speech-specific integration takes place here (Arnal *et al.*, 2009). More recently, it was shown that the auditory and visual responses in pSTS are linked in a meaningful way, whereby regions that respond to mouth motion, more strongly respond to vocal sounds and regions that respond to eye movements show no responses to vocal or non-vocal sounds (Zhu & Beauchamp, 2017). This suggests that the pSTS is a site where auditory and visual speech features are combined. Given that mouth movements almost always co-occur with hearing voices, it would be efficient for a neural population to encode these features together.

pSTS has also been shown to dynamically adjust the degree of connectivity between it and auditory and visual cortices depending on the reliability of the unimodal signals. Specifically, it was shown that when visual speech was blurred and the auditory speech clean, STS had greater functional connectivity with auditory cortex than visual cortex, and when the auditory speech was noisy and the visual speech was not blurred then STS increased connectivity with visual cortex and decreased connectivity with auditory cortex (Nath & Beauchamp, 2011). This suggests that STS up-weights the information coming from the more reliable stimulus modality – which agrees well with behavioral studies showing that people weigh perceptual judgements based on the reliability of the sensory modality (Ernst & Banks, 2002; Alais & Burr, 2004; Ma *et al.*, 2009).

Taken together, the research to date points towards a specificity in pSTS for auditory and visual speech feature encoding whereby this neural population can flexibly use the two independent forms of information to best represent the audiovisual speech signal and maximize comprehension.

The evidence for early multisensory integration occurring in core sensory cortices and a later stage of integration occurring in auditory association cortex has been formed into a multistage integration framework whereby it is proposed that audio and visual speech stimuli are integrated at multiple stages of processing (Fig. 2.11; Peelle & Sommers, 2015).

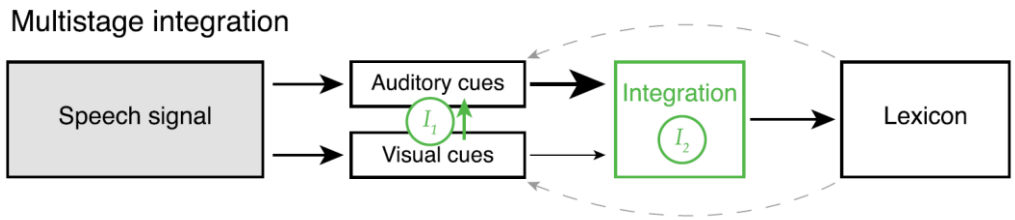


Figure 2.11. The multistage model of audiovisual speech integration. The schematic depicts how audiovisual speech consists of auditory and visual cues which are initially processed in separate sensory cortices. Already at the primary cortical stage of processing there is crossmodal interactions whereby visual inputs directly project to core auditory regions and impact low-level spectrotemporal processing. Then at a later stage, after both sensory inputs have gone through a degree of processing the visual and auditory inputs are integrated in order to constrain lexical selection. Adapted from (Peelle & Sommers, 2015)

2.3.3. The interaction of multisensory integration and attention

Several studies examining the integration of sensory signals using ERPs found very early (~40-50ms) multisensory interactions (Giard & Peronnet, 1999; Foxe *et al.*, 2000; Molholm *et al.*, 2002; and for review see: Schroeder & Foxe, 2005). This, and other work, led to suggestions that MSI is an automatic process that occurs without application of attention (Driver, 1996; Bertelson *et al.*, 2000). However, the function of integrating sources of information is to enhance perception, which is also the function of directed attention. Therefore, it is reasonable to suggest that these two processes would interact. Indeed, more recent work has demonstrated that when the demands on the cognitive processing system are high, MSI does not occur automatically (Alsius *et al.*, 2005; Alsius *et al.*, 2007; Morís Fernández *et al.*, 2015). This was further supported by evidence that attention to both visual and auditory modalities are necessary in order to observe a MS interaction (Talsma *et al.*, 2007), and when attention is directed elsewhere (i.e., attending something other than the auditory or visual signal) the MS interaction is actually reversed (Fig. 2.12; Talsma *et al.*, 2007). Therefore, there is clear evidence to suggest that MSI and attention interact in a manner that is dependent on environment and task demands.

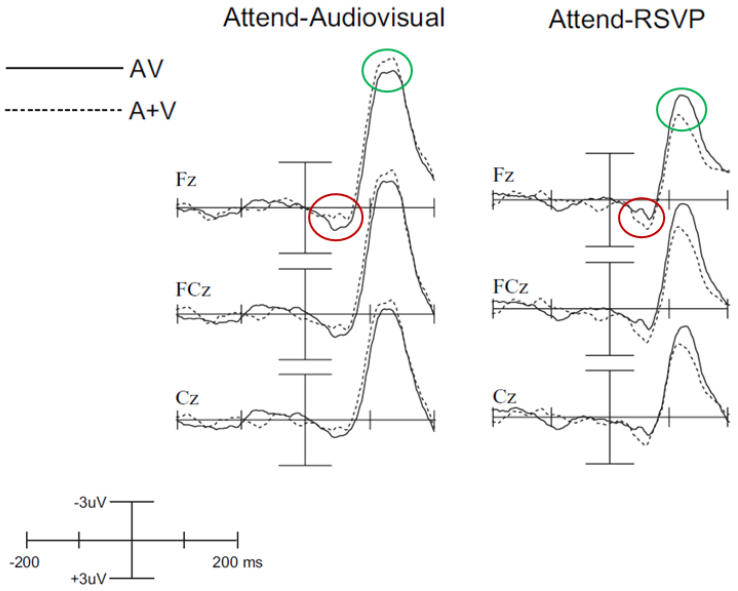


Figure 2.12. Left: Attention to both visual and auditory modalities are necessary in order to observe a MS interaction. The locations on the ERPs where AV differs from A+V are highlighted with red and green circles. Right: When attention is directed elsewhere (i.e., attending a rapid presentation of letters instead of attending the auditory sound or visual flash) the MS interaction is actually reversed. This is shown for 3 different locations on the scalp (Frontal electrode (Fz), fronto-central electrode (FCz) and central electrode (Cz)). Adapted from (Talsma et al., 2007)

2.4. Neurophysiology analysis methods

There are several different neuroimaging methods that are used to study the cortical dynamics of speech perception and multisensory integration in humans. This section focuses specifically on electroencephalography, the imaging method that was employed in this work, and the following sections describe in detail the analyses approaches used in this thesis.

2.4.1. Electroencephalography

Electroencephalography (EEG) is a non-invasive method that measures the neural activity on the scalp using electrodes. EEG measures the electrical field that is caused by synchronous activation of large groups of cortical neurons. Pyramidal cells are the most common type of cortical neuron and they are mostly oriented perpendicular to the scalp surface (Bok, 1959). When a large number of these neurons (~10^7-10^9) are activated synchronously it generates an electrical field which is measurable on the scalp surface. In order for the electrical activations to reach the scalp surface they must first pass through many layers of tissue such as the meninges covering the brain, the cerebrospinal fluid, the skull and the scalp, which all serve to attenuate the electrical signal. These layers of tissue also smear the signal across the scalp. Furthermore, the electrical signal on the scalp is

quiet small - in the order of micro volts, and so amplification is required to detect the signal. However, this also results in amplification of unwanted signals such as muscle activations that are of greater magnitude than the scalp recorded electrical activity. Together, these factors give rise to EEG having a poor spatial resolution as well as a low signal-to-noise ratio. The main advantage of EEG however, is its ability to capture fast neural dynamics – this makes it ideal for studying neural responses to rapidly changing stimuli such as speech.

2.4.2. Event-Related analysis

In order to extract some meaningful signal of interest from the EEG activity, event-based analysis has traditionally been used and is still popular today. This analysis involves taking each segment of the EEG time-locked to the onset of each stimulus presentation and averaging across the presentation. The resulting neural response is called the event-related potential (ERP) (Handy, 2005).

The reasoning behind this analysis is that you present the stimulus of interest multiple times spaced at some reasonable intervals such that there is enough time for the brain activity to come back to resting level between each stimulus presentation. Assuming that the response of interest to the stimulus is consistent across repetitions and the other 'noise' (i.e., spontaneous neural activity, muscle activity etc.) is truly random, uncorrelated with the stimulus response and has zero mean, then the signal of interest can be isolated by averaging.

This analysis framework has been the foundation of EEG analysis for decades and has provided insights into mechanisms of neural processing of audiovisual stimuli. However, the downside to this approach is that it requires multiple (possibly 100's) presentations of the same discrete stimulus in order to obtain a 'clean' time-locked average response, since the noise is often orders of magnitude greater than the signal of interest (Fig 2.13 shows ERPs plotted in response to auditory, visual and audiovisual stimuli). The need for using discrete stimuli also limits the possible experiments that can be performed, since one cannot use continuous stimuli such as streams of connected speech. Without the ability to study the brain using natural continuous speech, we are limited in our understanding of speech processing and so alternative methods and analysis techniques are required in order to progress the field on the cortical dynamics of speech processing (Hamilton & Huth, 2018).

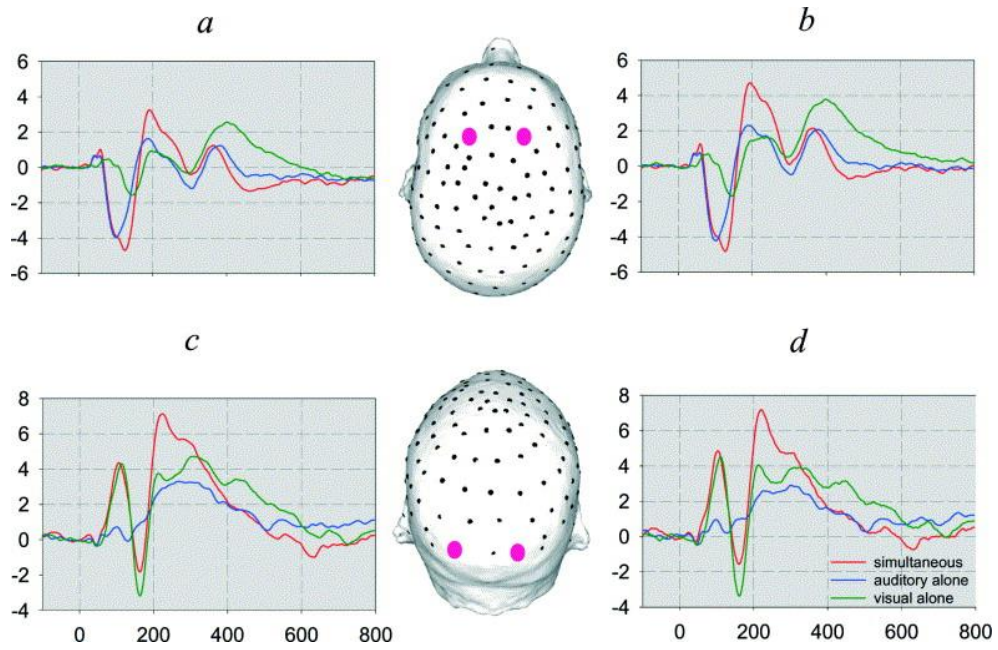


Figure 2.13. ERPs extracted from responses to auditory, visual and audiovisual stimuli at different scalp locations. The summed auditory and visual ERP is compared with the AV ERP to assess for multisensory interactions. Adapted from (Molholm et al., 2002).

2.4.3. Model-based TRF analysis

An alternative approach which has recently been proposed to overcome the shortcomings of the ERP based analysis is linear regression. This allows one to relate a continuous stimulus representation to the continuously changing neural response (Lalor & Foxe, 2010; Crosse *et al.*, 2016a). The method assumes that the relationship between the stimulus (input) and corresponding neural response measured by EEG (output) can be described by a linear time-invariant (LTI) system. However, the neural responses likely violate this assumption since the human brain is highly non-linear and time-variant by nature. Nevertheless, approximating it to a LTI enables us to use continuous and unrepeated stimuli in experimental paradigms as well as to model the brain-response relationship by an impulse response (which is referred to here as a temporal response function (TRF)).

Determining the impulse response of the system then allows us to estimate the output of the given system to a known input – this is referred to as forward mapping or response prediction. In a similar manner, the impulse response can be combined with a known output in order to reconstruct the stimulus – this is referred to as backward mapping or stimulus reconstruction. In this work, the stimulus representation is usually some characterisation of the speech sound (e.g., the sound envelope, the spectrogram etc.,) and the response is the recorded EEG.

2.4.4. Forward TRF mapping

In the forward mapping case, each EEG channel's activity can be independently estimated by training a model on a stimulus representation (e.g., the acoustic envelope) and the corresponding EEG response. These models are often referred to as encoding models because they describe how the system encodes information (Haufe *et al.*, 2014). The encoder, $w(\tau, n)$ represents the linear mapping from the stimulus feature time-series $s(t)$ to the multi-channel EEG response $\hat{r}(t, n)$ and is given by the following equation:

$$\hat{r}(t,n) = \sum_{\tau} w(\tau, n) s(t - \tau)$$

where $t = 1 \ldots T$ is the time sample, $\tau = \tau_{min} \ldots \tau_{max}$ is the relative time lag in samples, and $n = 1..N$ is the number of EEG channels. The model w can be solved by ridge regression (Tikhonov & Arsenin, 1977)

$$w = (S^T S + \lambda I)^{-1} S^T r$$

where λ is the ridge regression parameter, I is the identity matrix and the matrix S is lagged time series of the stimulus feature:

$$S = \begin{bmatrix} s(1-\tau_{min}) & s(-\tau_{min}) & \cdots & s(1) & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots & s(1) & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots & \vdots & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \cdots & s(1) \\ s(T) & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ 0 & s(T) & \cdots & \vdots & \vdots & \cdots & \vdots \\ \vdots & 0 & \cdots & \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & s(T) & s(T-1) & \cdots & s(T-\tau_{max}) \end{bmatrix}$$

Time-lags are included in the stimulus representation because the neural response is not instantaneous but instead evolves over hundreds of milliseconds, and the inclusion of multiple time-lags allows us to capture the change in response over time. Negative time-lags correspond to the EEG leading the stimulus and positive time-lags correspond to the EEG lagging the stimulus. The resulting model therefore is 2-dimensional – it contains weights for each EEG channel at each time-lag and can be interpreted in a similar way as an ERP. Leave-one-out cross-validation is used for model estimation in order to prevent overfitting. The quality of the model fit is assessed by evaluating the correlation between the predicted and the original EEG signal using Pearson's r.

One advantage of the forward mapping approach is that the stimulus representation can contain multiple features. This is useful because it allows us to study

the encoding of a variety of speech features such as a time by frequency representation (spectrogram) or a representation based on the phonetic features (Di Liberto *et al.*, 2015) in the speech or even the visual speech features (O'Sullivan *et al.*, 2017). However, a disadvantage of this approach is that each neural channel is predicted independently and so the model cannot integrate information across channels.

2.4.5. Backward TRF mapping

The forward encoding necessarily requires the independent prediction of each EEG channel. However, the scalp EEG channels have a relatively high correlation with each other since nearby channels capture similar underlying brain areas when they become activated. Therefore it would be beneficial if we could model all EEG channels simultaneously in order to study the neural representation of the given stimulus. Backward direction TRF mapping or stimulus reconstruction allows this, since it combines all EEG channels across multiple time-lags in order to reconstruct an estimate of the stimulus. The model, termed a decoder, $g(\tau, n)$ represents the linear mapping from the continuous multi-channel EEG response $r(t, n)$ back to the stimulus time-series $\hat{s}(t)$.

$$\hat{s}(t) = \sum_{n=1}^{N} \sum_{\tau} g(\tau, n) r(t + \tau, n)$$

where $t = 1 \ldots T$ is the time sample, $\tau = \tau_{min} \ldots \tau_{max}$ is the relative time lag in samples, and $n = 1..N$ is the number of EEG channels.

As in the forward mapping, the model g can be solved by performing ridge regression:

$$g = (R^T R + \lambda I)^{-1} R^T s$$

where λ is the ridge regression parameter, I is the identity matrix and the matrix R is lagged time series of the EEG data:

$$R = \begin{bmatrix} r(1-\tau_{min},1) & r(-\tau_{min},1) & \cdots & r(1,1) & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots & r(1,1) & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots & \vdots & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \cdots & r(1,1) \\ r(T,1) & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ 0 & r(T,1) & \cdots & \vdots & \vdots & \cdots & \vdots \\ \vdots & 0 & \cdots & \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & r(T,1) & r(T-1,1) & \cdots & r(T-\tau_{max},1) \end{bmatrix}$$

A similar fitting procedure is performed as was in the forward mapping case using leave-one-out cross-validation and evaluating performance by correlating the reconstructed

estimate of the stimulus with the true stimulus.

Since the backward model integrates across EEG channels and time-lags it has the ability to be more sensitive than other EEG analyses such as the forward modelling approach or time-averaging methods. The backward model weights however, are not readily interpretable since they are sensitive to noise and need to be converted to forward activation patterns in order to be physiologically interpretable (Haufe *et al.*, 2014). The main drawback of this approach is that it is currently only possible to reconstruct univariate stimulus representations and so is limited in its capacity to study different speech representations (e.g., the speech spectrogram or phonemes).

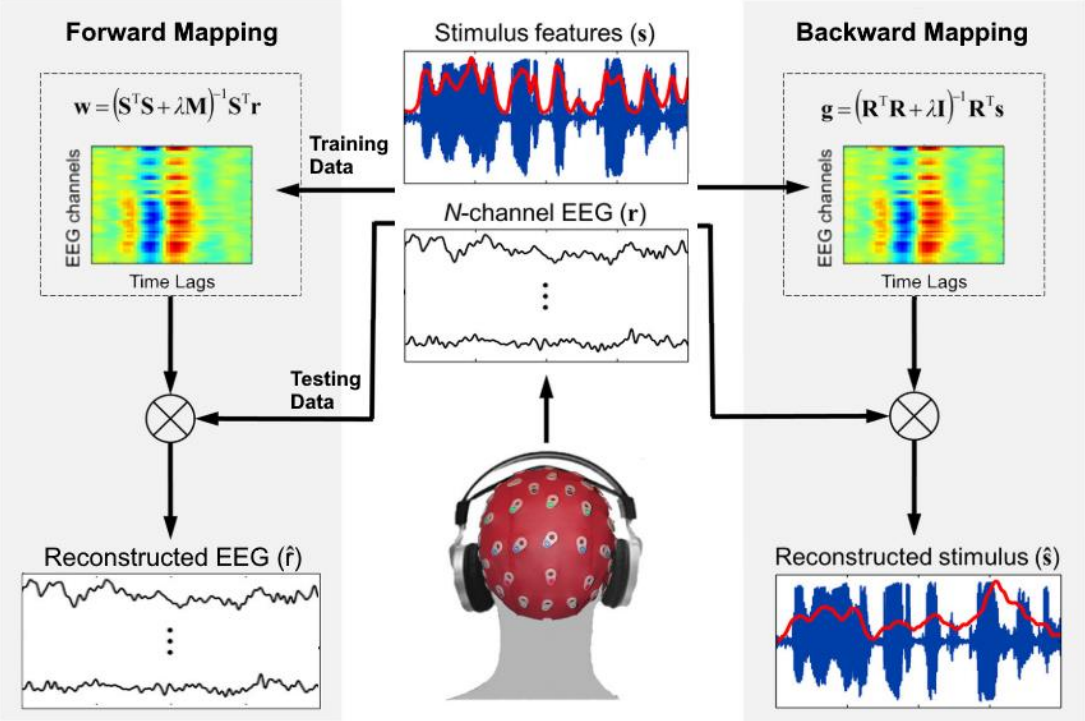


Figure 2.14. Schematic of forward and backward TRF mapping. Left: The forward mapping finds an encoding model that allows the prediction of new EEG data from a known stimulus time-series. Right: Backward mapping involves building a decoder which reconstructs a new, unseen stimulus time-series from the known EEG data. To prevent model-overfitting, ridge regression and leave-one-out cross-validation is used.

As previously mentioned, there are many ways in which we can characterise the speech signal, using univariate representations such as the envelope, or multivariate representations such as the spectrogram or phonetic feature representation. However, since these regression models in their general form allow only univariate-multivariate comparison, the forward encoding models can be used to study discrete and multidimensional representations of the speech like phonemes but they suffer from poor sensitivity due to the need to predict each neural channel independently. While the backward modelling approach can be more sensitive due to pooling data across neural

channels (e.g., EEG electrodes), it cannot, at present, be used with discrete and multidimensional representations of speech.

To overcome the limitation of relating multiple features to the multivariate EEG response we use a method known as canonical correlation analysis (CCA) which is described in the next section.

2.4.6. Canonical correlation analysis

Canonical correlation analysis is a method for relating two sets of variables which was proposed over 70 years ago (Hotelling, 1936). It has been applied in many scientific fields since then (for e.g., learning semantic relationships between images and their associated text: Hardoon *et al.*, 2004) and more recently, it has been applied to analyse neuroimaging data (Correa *et al.*, 2010; Varoquaux *et al.*, 2010; Bilenko & Gallant, 2016; de Cheveigne *et al.*, 2018; Lettieri *et al.*, 2019; Rampinini *et al.*, 2019; Donhauser & Baillet, 2020).

CCA works by transforming two given sets of multidimensional data into a space in which they are maximally correlated. This linear transformation is based on finding a set of basis vectors for each of the given data sets such that the correlation between the variables, when they are projected on these basis vectors, are mutually maximised. This has the potential to reveal hidden relationships between the variables since the correlation between variables is dependent on the co-ordinate system in which the variables are described. The optimisation problem for CCA is formulated as a generalised eigenproblem.

Given the data matrix $s(t)$ of size $T \times J_1$, where in the context of the work presented here, $s(t)$ can be thought of as the multidimensional stimulus representation, where T is time and J_1 is the number of features in that representation (e.g., $J_1 = 16$ frequency bands of a spectrogram, or $J_1 = 19$ phonetic features etc.), and the second data matrix $r(t)$ of size $T \times J_2$, with $r(t)$ representing the corresponding EEG data where T is time and $J_2 = n\ x\ \tau$, where n is number of EEG channels and τ is the number of time-lags used. For these two data matrices, CCA produces transform matrices $\bar{A}_1$ and $\bar{A}_2$ of sizes $J_1 \times J_0$ and $J_2 \times J_0$, where J_0 is at most equal to the smaller of J_1 and J_2.

These transform matrices ($\bar{A}_1$ and $\bar{A}_2$) are found by performing eigendecomposition on the joint (stimulus and response) covariance matrix. The first pair of canonical components (CC) define the linear combinations of each data set with the highest possible correlation. The next pair of CCs are the most highly correlated

combinations orthogonal to the first, and so-on (de Cheveigne *et al.*, 2018).

$$\sum_{1}^{J1} s_{J1}(t)\bar{A}_1 \rightarrow CC_{min(J_1,J_2)} \rightarrow \rho \leftarrow CC_{min(J_1,J_2)} \leftarrow \sum_{1}^{J2} r_{J2}(t)\bar{A}_2$$

Where $s_{J1}(t)$ is the stimulus representation with $J1$ features, $\bar{A}_1$ are the weights applied to those features and $CC_{min(J_1,J_2)}$ are the resulting canonical components. On the other side of the equation, $r_{J2}(t)$ represents the EEG response with $J2$ features. In this case, the response feature matrix is described as EEG channels by time-lags, where 0-500ms were the time-lags used. $\bar{A}_2$ are the weightings applied to the response matrix and $CC_{min(J_1,J_2)}$ are the resulting canonical components.

In order to find the canonical components however, inversion of the sample correlation matrices is required which can be ill-posed (when there is a large degree of autocorrelation in the sample correlation matrices). To overcome this, one common solution is ridge regularization, which involves adding a multiple of the identity matrix to the sample correlation matrices (Tikhonov & Arsenin, 1977). As was the case for the backwards TRF modelling, the ridge parameter is applied to the covariance matrix of the time-lagged EEG data and leave-one-out cross-validation is used to prevent over-fitting (Vinod, 1976; Leurgans *et al.*, 1993; Cruz-Cano & Lee, 2014).

2.5. Summary

In this chapter we have discussed current theories on the neurophysiology of auditory speech perception. In particular we have presented the idea that auditory speech is processed in a hierarchical manner and that speech specific encoding is thought to take place outside of core auditory cortex, specifically in STS and STG. We then presented how visual speech information is related to auditory speech and how this information may be processed in visual cortex or in heteromodal cortex. The left posterior temporal cortex has been implicated as being sensitive to visual speech features and proposed as a region in which visual phonetic information is extracted and fed-forward to auditory speech processing regions.

In the next section we presented commonly used methods for measuring multisensory interaction effects and then discussed how STS has been repeatedly implicated as a region where multisensory integration occurs and more recent evidence has pointed to the posterior region of STS jointly encoding mouth movements and vocal sounds - demonstrating its importance for audiovisual speech integration. We then

presented a brief overview on the debate of the interaction between multisensory integration and attention.

In the final section of this chapter, we provided an introduction on the neuroimaging technique used in this thesis, electroencephalography (EEG). Next, we describe analysis methods applied to this type of data, in particular the traditionally used event-related (ERP) analysis and the more advanced model-based approaches based on linear regression which include both forward encoding models trained to predict EEG and backward reconstruction models trained to reconstruct the speech envelope. We also discuss canonical correlation analysis (CCA) as an alternative technique for data analysis which allows you to relate two multivariate data sets to each other. The latter techniques, namely linear regression and CCA are the analysis methods used throughout this work.

In the next chapter, we present the first study of this thesis which examines how cortical tracking of the speech signal is modulated by the relevance of the visual input in a multi-talker scenario.

Chapter 3.

Attention to visual speech at a cocktail party affects stimulus reconstruction and alpha power modulations of EEG

3.1. Introduction

In our daily social interactions we frequently solve what is termed 'the cocktail party problem' (Cherry, 1953). This refers to the brain's ability to segregate and selectively attend to a single source of speech while suppressing all others. Critical insights into the neural underpinnings of selective attention to speech have been provided by imaging techniques such as magnetoencephalography (MEG) and electrocorticography (ECoG). In particular, studies employing these techniques have characterized the internal representations of attended and unattended speech streams at multiple levels along the auditory cortical hierarchy (Ding & Simon, 2012b; Ding & Simon, 2012a; Mesgarani & Chang, 2012; Golumbic *et al.*, 2013a; Puvvada & Simon, 2017).

The vast majority of studies however, do not take into account the fact that our sensory input in everyday environments is almost exclusively multimodal - containing sources of both auditory and visual information (but see Golumbic *et al.*, 2013b). This has a number of implications regarding the processing of attended and unattended speech streams. Firstly, visual speech (lipreading) is known to aid speech comprehension (Reisberg *et al.*, 1987; Crosse *et al.*, 2015a), a benefit which is even more evident in sub-optimal listening conditions (Sumby & Pollack, 1954; Grant & Seitz, 2000; Ross *et al.*, 2007; Crosse *et al.*, 2016b). In addition, since the visible articulatory movements of the speakers face are temporally correlated with the acoustic energy of their speech (Summerfield, 1992; Grant & Seitz, 2000; Chandrasekaran *et al.*, 2009), inclusion of visual cortical activity related to these movements improves the fidelity with which one can relate EEG signals to audio speech (Crosse *et al.*, 2015a). Furthermore, audio speech tracking is enhanced in auditory cortex in the presence of congruent visual speech (Schroeder *et al.*, 2008; Crosse *et al.*, 2015a; Peelle & Sommers, 2015; Crosse *et al.*, 2016b) and brain regions specialized for processing multisensory inputs will also influence auditory processing and perception (Calvert *et al.*, 2000; Sekiyama *et al.*, 2003; Wright *et al.*, 2003; Nath & Beauchamp, 2011; Zhu & Beauchamp, 2017).

Furthermore, there is little known about how speech tracking is influenced when the speaker at which the listener is looking, is not the one whom they are attending (i.e., eavesdropping). In this case, the visual information is irrelevant for attending to the desired speech stream, and may have a detrimental impact on selective attention, and it is unknown how neural correlates of speech tracking are affected under such circumstances. This type of environment also presents an opportunity to explore neural measures specifically related to attention to visual speech.

One well-established marker of visual attention is alpha power (Pfurtscheller & Klimesch, 1991) which could potentially be used to index the relevance of the visual information that the subject is looking at (Kelly *et al.*, 2006). Specifically, we might expect to see a relative increase in scalp alpha power when the visual information is distracting in the eavesdropping condition compared to the case where the visual information actually benefits speech comprehension.

These considerations are important for understanding how the relevance of visual speech impacts auditory speech processing and the application of selective attention in more natural environments. We investigate this question here by using stimulus reconstruction to relate the EEG signal to the speech envelope of 2 different speakers: one for which the visual speech is relevant and the other whereby the visual speech is irrelevant. It is expected that differences in speech tracking between these two conditions will influence our ability to reconstruct the speech envelope.

The results of this study were published as a research article:

O'Sullivan, A. E., Lim, C. Y., & Lalor, E. C. (2019). "Look at me when I'm talking to you: Selective attention at a multisensory cocktail party can be decoded using stimulus reconstruction and alpha power modulations." *European Journal of Neuroscience*, 00: 1–14. https://doi.org/10.1111/ejn.14425.

3.2. Material and Methods

Participants. Seventeen native English speakers (eight males; age range 18-30 years) took part in the experiment. All participants were right-handed, were free of neurological diseases, had self-reported normal hearing, and had normal or corrected-to-normal vision. Informed consent was obtained from all participants before the experiment, and they received monetary compensation for their time. The study was approved by the Research Subjects Review Board at the University of Rochester.

Procedure. In this experiment subjects were presented with two competing speakers via headphones, one speaker was presented at 0° (no HRTF applied - A1) with the accompanying video of the speaker displayed on the monitor, which was also at 0° (V1), while the other speaker was presented in the auditory modality at 30° to the right (referred to as A2). The content consisted of two audiobooks of classic works of fiction (story 1: Journey to the Centre of the Earth and story 2: Twenty Thousand Leagues under the Sea by Jules Verne) read by two different male speakers.

There were two experimental conditions, which differed in the relevance of visual information: (1) AVc (audiovisual congruent) - subjects were instructed to attend to the audiovisual speaker and ignore the competing audio-only speaker (A2). See attend AVc; Fig. 3.1, left.

(2) AVi (audio-visual incongruent) - subjects were asked to attend to the audio-only speaker (A2), while looking at the screen showing the (in this case incongruent) video of the other speaker talking (V1) and ignoring the A1 audio (attend AVi; Fig. 3.1, right). It is important to point out that although the subject is attending audio-only speech in this case, in order to do so the subject must suppress the competing AVc speech. This type of task is more closely related to attending incongruent audio-only speech and so the condition is named accordingly.

There were 20 trials per condition, each trial lasting 1 minute, and each trial began where the story ended on the previous trial. After every trial subjects were required to answer between 4 and 6 multiple-choice questions on both stories. Each question had 4 possible answers. One caveat here is that the order of presentation of the stories was not counterbalanced across conditions and so it is conceivable that one part of the story may be easier to attend than another. Subjects were instructed to look at the speaker's eyes for the duration of the trial in both conditions, and to minimize eye blinking and all other motor activities.

Stimuli.

Audiovisual stimuli: The first audiobook (Journey to the Centre of the Earth) was used for the audiovisual speech stimuli. This consisted of recordings of a hired actor reading from a prompter into the camera. This actor was not known to be familiar to any of the participants. Subtitles were displayed on a screen adjacent to the camera and aligned with the centre of the camera lens to prompt the actor. Recordings were carried out in a soundproof booth using a Nikon D750 camera and a ZOOM H4nPro audio recorder. The videos consisted of the speaker's head and shoulders with a black background (see

screenshot in figure 3.1). The audio and video files were then combined in VirtualDub video editor (synchronised using a clap) and rendered into videos with a frame rate of 30 fps using the Xvid MPEG-4 Codec video encoder. The videos were then cropped to a frame size of 1280x1080 so that the face of the speaker was centred in the frame for all videos. Forty combined audiovisual files were then clipped to be ~1 minute in length. The visual stimuli were presented on a 35 inch LCD monitor (ASUS Predator), operating at a refresh rate of 60 Hz, and participants were seated at a distance of 70 cm from the display. The face of the speaker subtended a visual angle of 16°.

Audio-only stimuli: The second audiobook (Twenty Thousand Leagues under the Sea) was used for the content of the audio-only stimuli. The reader of this audiobook was also not known to be familiar to any of the participants. The audio files were convolved with a HRTF for a sound presented at 30° to the right (taken from the CIPIC Database: Algazi *et al.*, 2001) in MATLAB. The intensity of each soundtrack (from both stories) measured by root mean square, was normalized in MATLAB. The audio file used as the competing story was manually aligned with the audio file taken from the video so that the two speakers began talking at the same time for each trial and the aligned files were then summed together in MATLAB.

Audio stimuli were presented diotically through Sennheiser HD650 headphones at a comfortable level. Stimulus presentation and data recording took place in a dark sound attenuated room and stimulus presentation was controlled using Presentation software (Neurobehavioral Systems). The stimuli used in this study will be made available upon reasonable request to the author or advisor.

Eye tracking. Participant's eye movements were also monitored throughout the experiment to ensure that they followed the instructions given. Eye tracking was recorded for the duration of each trial using the EyeLink 1000 infrared eye tracker (SR Research, Ottawa, Canada). A chin rest was used to stabilise the participants head during trials. The eye tracker was desktop mounted and used in monocular mode with a sampling rate of 500 Hz. Calibration and verification was carried out at the beginning of each condition using 9 targets distributed over the entire screen. In order to compare the duration of time a subject spent looking at the eyes versus the mouth of the speaker separate regions of interest (ROI) were created for each trial. The ROIs consisted of rectangular boxes which captured the left eye, right eye, and the mouth of the speaker. The ROIs were large enough to capture the entire mouth during opening as well as account for any head movements the actor made while speaking. Importantly, the ROIs were the same size for each trial

(i.e., contained the same number of pixels) and were constructed such that the sum of the area in the eye ROIs equalled the area of the mouth ROI (as in Gurler *et al.*, 2015). Eye movements were recorded for 15 out of the 17 subjects due to technical issues, and so the data presented here are from 15 subjects.

EEG data preprocessing. EEG data were acquired using an ActiveTwo system (BioSemi, The Netherlands) from 128 scalp electrodes and two mastoid electrodes at a sampling rate of 512 Hz. Triggers indicating the start of each trial were recorded along with the EEG. Subsequent preprocessing was conducted off-line in MATLAB; the data were bandpass filtered between 0.3 and 30 Hz, downsampled to 64 Hz, and referenced to the average of the mastoid channels. EEG channels with excessive noise were detected using EEGLAB channel rejection methods (Delorme & Makeig, 2004). Specifically, we used the spectrogram, kurtosis and probability methods provided by the EEGLAB toolbox to determine an excessively noisy channel. Channels were also classified as noisy if the variance was 3 times greater than the average variance of all channels. Noisy channels were replaced by spline-interpolating the surrounding clean channels in EEGLAB (Delorme & Makeig, 2004).

Data analysis. To relate the speech stimulus to the EEG, the acoustic envelope of the speech signal was extracted. This was done by bandpass filtering the signal into 128 logarithmically-spaced frequency bands between 100 and 6500 Hz using a gammatone filterbank (Irino & Patterson, 2006). Afterwards, the envelope at each of the 128 frequency bands was extracted by taking the absolute value of the Hilbert transform, and the broadband envelope was calculated by averaging over the 128 narrowband envelopes.

Our objective was to determine from the EEG signal which speaker a participant was attending in each trial. To achieve this, we trained a linear model (decoder) to reconstruct an estimate of the attended envelope from the EEG for each trial. If the EEG used to train the decoder is from an attend congruent AV (AVc) speech trial, then the resulting decoder is referred to as 'AVc Decoder'. If the EEG comes from a trial where subjects attend the audio-only speech (eavesdropping), the decoder is referred to as 'AVi Decoder'. The linear model used to estimate the speech envelope, $\hat{s}(t)$ from the EEG data, $r(t)$ is implemented as follows (Crosse *et al.*, 2016a):

$$\hat{s}(t) = \sum_{n=1}^{128} \sum_{\tau=0}^{500ms} r(t+\tau, n)\, g(\tau, n)$$

Where $\hat{s}(t)$ is the envelope reconstruction, $r(t+\tau, n)$, is the EEG response at channel n

and time lag τ, and $g(\tau, n)$ is the linear decoder for the corresponding channel and time lag. The decoder integrates EEG from 0 ms to 500 ms poststimulus to reconstruct each sample of the speech envelope. This ensures that we capture the relevant temporal information in the EEG that relates to each sample of the stimulus that we are trying to reconstruct. Leave-one-out cross-validation was used to determine the performance measure of our stimulus reconstruction and the linear decoder was optimized for each condition by maximizing the correlation between the attended and reconstructed envelopes, $s(t)$ and $\hat{s}(t)$ respectively.

Decoding attention

1. Based on stimulus reconstruction: To test the ability of a decoder to determine the attended speaker we take the reconstructed speech envelope output from the decoder and evaluate the reconstruction accuracy by determining a correlation coefficient (Pearson's r) between the reconstructed speech envelope and the actual attended and unattended speech envelopes, which we will refer to as r-attended and r-unattended, respectively. Thus, we would consider a trial to be correctly decoded if the reconstruction had a greater correlation with the attended speech envelope (i.e., if r-attended > r-unattended). The percentage of trials where we correctly decoded attentional-selection will hereafter be referred to as decoding-accuracy. However, when trying to decode EEG data to determine who someone is attending to, we do not know in advance whether they are attending to the AVc or AVi stimulus. As such, it is not clear whether it would be optimal to use an AVc decoder or an AVi decoder.

To test if the stimulus reconstruction method is sensitive enough to discriminate between the attended speech modalities (i.e., AVc vs AVi) we can use the output correlations (attended and unattended) from both decoders. This gives us four reconstruction values (r-attended and r-unattended for each of the two decoders). We then classify the trial as AVc or AVi using a linear discriminant analysis (LDA) classifier trained on all remaining trials using leave-one-out cross-validation.

2. Based on alpha power: It may also be appropriate to use additional neural measures for determining the attended modality (AVc vs AVi). In particular, alpha power has been linked to an active mechanism for tracking attended auditory speech (Wöstmann *et al.*, 2016), as well as for suppressing visual attention (Kelly *et al.*, 2006; Romei *et al.*, 2008; for review see: Foxe & Snyder, 2011). Therefore the distribution of alpha power across the scalp may allow us to discriminate between the attended speech modalities presented here. To test this, we bandpass filtered the EEG data between 8-12 Hz and

calculated the absolute value of the Hilbert transform at every channel. We then took the average of the alpha band envelope over the duration of each trial. Then we had 20 trials for attend AVi and 20 trials for attend AVc which we sought to classify using LDA.

Statistical analyses. To statistically assess whether classification performance is above the theoretical chance level, we used a non-parametric permutation test (Combrisson & Jerbi, 2015). First, to establish a null distribution of the classification accuracies, we repeated the LDA 1000-times with randomly permuted classification labels. Afterwards, we used the tail of this empirical distribution to calculate the p-value for the original classification. This was done for each subject and each condition separately. All group level statistical comparisons were conducted using two-sided Wilcoxon signed-rank tests and a significance threshold of 1% was chosen where multiple comparisons are made. All numerical values are reported as mean ± SD.

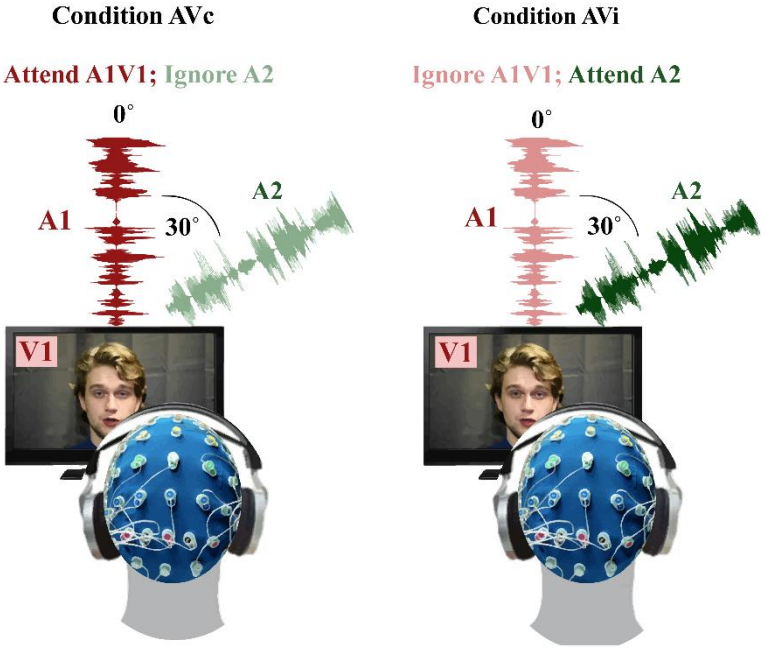


Figure 3.1 Experiment setup. In this experiment, subjects performed a multisensory version of the traditional cocktail party. In one condition, they attended to congruent A1V1 speech while ignoring competing A2 speech (left panel). In the second condition, they attended A2 speech in the presence of competing A1V1 speech (right panel). Linear regression was performed between the EEG and envelope to determine the attended speech stream.

3.3. Results

Behaviour: speech comprehension. All subjects performed better at answering questions on the attended story compared with the unattended story in both conditions (AVi: $p=2.9\text{x}10^{-4}$, AVc: $p=2.9\text{x}10^{-4}$; Fig. 3.2). This demonstrates that subjects were successful at attending the requested story in both conditions. However, subjects

performed significantly better at answering questions while attending AVc speech compared with attending AVi (AVc: 76.1 ± 9.7%, AVi: 61.4 ± 12.8%; $p=2.9x10^{-4}$; Fig. 3.2).

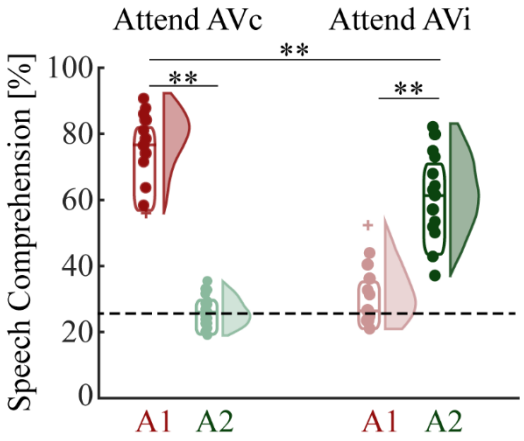


*Figure 3.2 Speech comprehension. Subject's performance answering questions on the attended and unattended stories. The dashed line indicates chance level performance (25%). Left: Performance for attend AVc condition (attend A1V1). Right: Performance for attend AVi condition (attend A2). Comparison of performance in both conditions revealed that AVc performance was significantly better than AVi. ** p < 0.01.*

Envelope reconstruction. We examined the effect of the relevance of the visual input on cortical tracking of the attended envelope. In doing so, we tested the performance of the AVc and AVi decoders and expected to find differences in how well they could reconstruct the envelope of the attended speech modality. Indeed, when subjects attended AVc (A1V1) speech we found a larger reconstruction accuracy for the A1 envelope in comparison with the unattended A2 envelope ($p=2.9x10^{-4}$; Fig. 3.3A, left). Similarly, when subjects attended to AVi speech (A2V1), we found the reconstruction accuracy for the A2 envelope was significantly higher than the unattended A1 envelope ($p=3.5x10^{-4}$; Fig. 3.3A, right). However, the difference between the attended and unattended reconstruction accuracies is not as large when attending AVi speech compared with attending AVc speech ($p=2.9x10^{-4}$). This suggests that attentional modulation of speech tracking is more effective for attending AVc than AVi speech and is in agreement with the behavioural results. We compared the speech comprehension accuracy and envelope reconstruction accuracy and found no significant correlation across subjects (AVi: r=0.23, p=0.37; AVc: r=0.29, p=0.25). At the single trial level we found no correlation between speech comprehension and reconstruction accuracy for attend AVi (r=0.03, p=0.57), however, we found a weak positive correlation between the comprehension accuracy for attend AVc and the envelope reconstruction accuracy (r=0.14, p=0.01).

We then wanted to investigate if the two decoders (AVc and AVi) were similar to each other given the observed similarity in the activation patterns (Fig. 3.3E). To test this, we

first investigated if there are differences in reconstruction accuracy for the AVc and AVi decoders across conditions. We used the decoder trained on one condition to reconstruct the attended envelope in the other condition. We find that the AVc decoder reconstructs the attended AVi envelope with significantly greater accuracy than the unattended envelope (AVc decoder on AVi EEG: p=0.003; Fig. 3.3B, left). In a similar manner, we tested the AVi decoders ability to reconstruct the envelope from EEG data of subjects attending AVc speech and find it is significantly better at reconstructing the attended AVc envelope relative to the ignored envelope (AVi decoder on AVc EEG: $p=2.9x10^{-4}$; Fig. 3.3B, right). This shows that both decoders better reconstruct the attended speech envelope at the trial average level in both conditions. In the next section, we compare the decoder's performances at the level of a single trial.

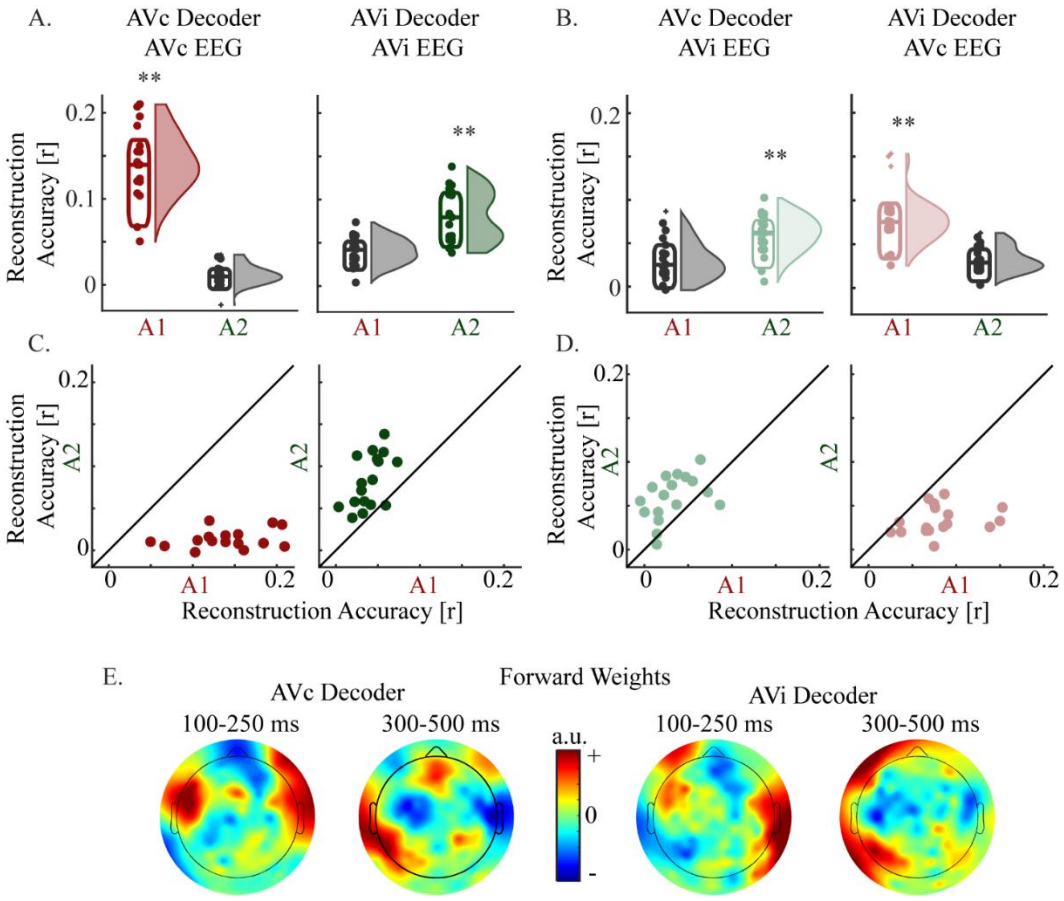


Figure 3.3. Speech envelope reconstruction. (A-D) Top: Speech envelope reconstruction accuracies for each decoder and condition. Bottom: Single subject reconstruction accuracies for attended and unattended envelopes (trial averaged). E. Transformed scalp decoder weights for AVc and AVi decoders at specific time lags. Backward weights are transformed into forward weights for interpretability (Haufe et al., 2014).

Decoding selective attention. Based on stimulus reconstruction: The results presented above are based on envelope reconstruction accuracies that have been averaged across trials. However, we are also interested in the performance of the decoder at determining

the attended speech at the level of a single (1 minute) trial. The traditional approach for determining the attended speech stream has been to take the generated envelope reconstruction and compare it with the heard speech envelopes (by calculating Pearson's r), then the envelope which has the largest correlation with the reconstructed envelope is deemed the attended envelope (O'Sullivan *et al.*, 2014). This measure has worked quite well, providing reasonably high accuracy (>85%) for single trial (1 minute) EEG data, and is easy to compute (O'Sullivan *et al.*, 2014). Applying the same approach here, we find that the AVc decoder decodes attention to AVc speech with an accuracy of 94.4 ± 6.3% (Fig. 3.4A, left), and the AVi decoder decodes attention to AVi speech with an accuracy of 74.4 ± 16.7% (Fig. 3.4A, right), these are both above the 5% significance level of 70% (binomial test).

We then examined how well the decoders would work at decoding attention to speech in the other modality. This was done to quantify the similarity of the two decoders. For the AVc decoder we found it to be significantly worse when used to decode attention to AVi speech ($p=0.01$), with an accuracy of 65 ± 15.1% (a decrease of 9.4%). Similarly, the AVi decoder performs poorly relative to the AVc decoder at decoding attention to AVc speech with an accuracy of 70.6 ± 15.4% ($p=6.1\text{x}10^{-5}$; Fig 3.4A), which corresponds to a decrease of 23.8%. Thus for both conditions cross-modal decoding is significantly worse than modality-specific decoding, and this decrease in decoding performance is significantly larger for the AVc decoder ($p=0.01$).

Since the decoders are trained in a subject-specific manner, it is relevant to examine their performance at the single subject level. For attending to AVc speech, we see that the AVc decoder is better for almost all subjects (for 2 subjects it performs similarly to the AVi decoder, Fig. 3.4B, left). We see a somewhat similar pattern emerge for the attend AVi condition, as the AVi decoder performs better for all but 4 subjects (Fig. 3.4B, right). This shows that modality-specific decoders are optimal for reliably determining the attended speech.

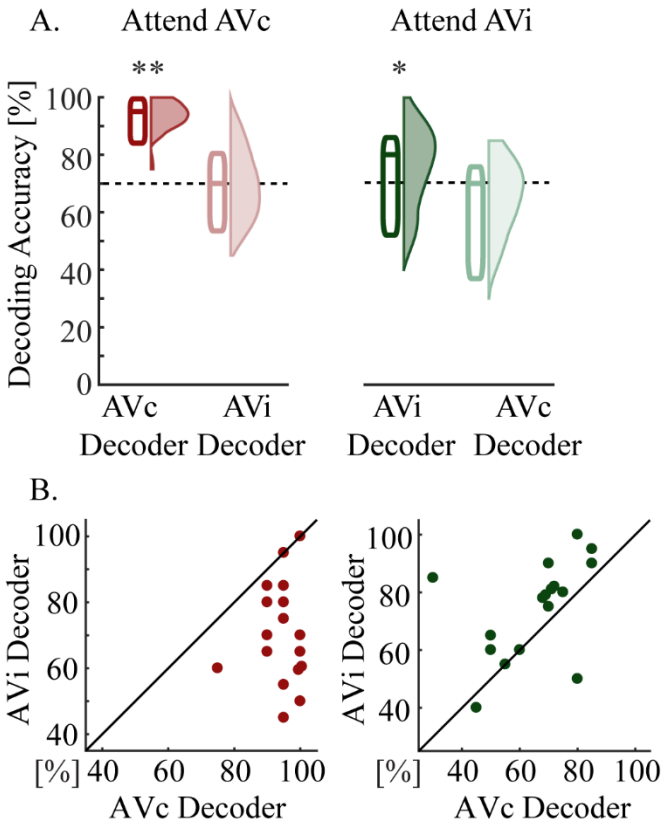


Figure 3.4 Decoding attention on single trials. A. Decoding performance for each condition and decoder. The dashed line indicates the 5% significance level determined using a binomial test. Decoding accuracy of AVc decoder for attend AVc is significantly larger than decoding attend AVc using the AVi decoder ($p=6.1\times10^{-5}$). Similarly, the AVi decoder performs significantly better than the AVc decoder at decoding attention to AVi speech ($p=0.01$). The single subject scatter plot is removed from this figure due to extensive overlap of points and instead is shown in b. B. Single subject decoding accuracy using same or cross modality decoders. For the majority of subjects in both conditions using the modality specific decoder works best.

Decoding attention when relevance of visual input is unknown. Our results so far have shown that it is optimal to use the decoder which has been trained on a particular speech modality (AVc or AVi) in order to decode the attended speech. However, in natural scenarios, we don't know whether the listener is attending the face they are looking at or eavesdropping. Therefore, we don't know which decoder (AVc or AVi) to use in order to achieve the best possible performance.

In order to overcome this, we used two methods: (1) LDA on reconstruction accuracies and (2) Alpha power.

(1) Based on stimulus reconstruction: We used LDA to combine the output from both decoders in order to determine the attended speech modality. Specifically, as the input to LDA, we use the output correlation values with the attended and unattended envelopes of both AVc and AVi decoders. These 4 correlation values are then used to train a linear classifier to determine which envelope is attended. This classifier then predicts which

speech modality was attended for a left out trial (done for every trial in both conditions). Using this method, we found that we could determine the attended speech with 75.6 ± 12.8% accuracy. The mean chance level classification was 49.9 ± 0.47% and was computed by averaging the permutation test classification accuracies across all trials and repetitions (Fig. 3.5A). Across subjects, the classification accuracies were significantly larger than the chance level ($p=2.9x10^{-4}$). We also ran the analysis at the single-subject level. For all subjects, the classification accuracies averaged across all trials were higher than the average classification accuracy obtained from the permutation test, and for twelve out of seventeen subjects, this difference was significant (Fig. 3.5B). For the remaining five that do not reach the significance threshold, three of them had an average performance above 60%, while the subject that reached only 50% accuracy had the poorest average behavioural performance with a comprehension score of 53%.

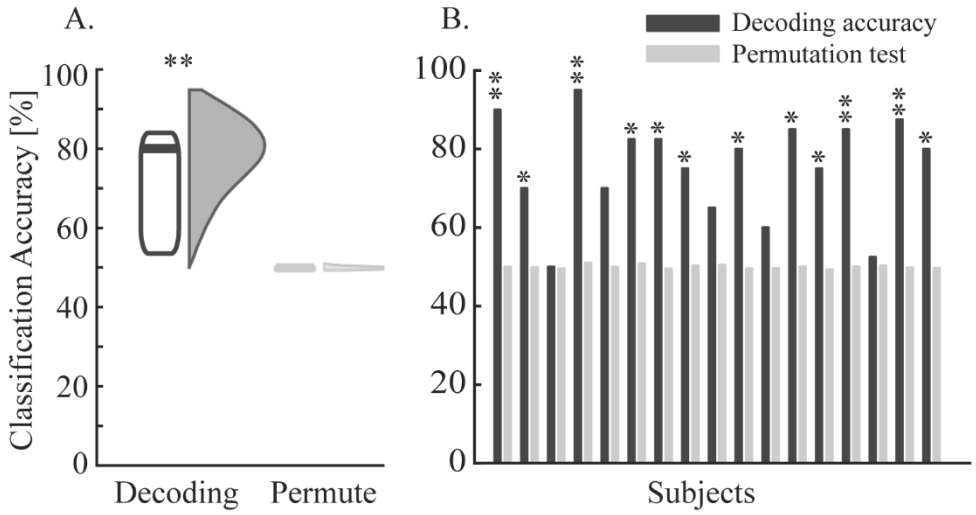


*Figure 3.5 Speech modality classification. Classification of attended speech modality using LDA which takes the reconstruction accuracies obtained from both decoders to determine the attended modality. A. Classification accuracy is significantly greater than chance level (determined by permutation test). B. Single subject classification scores are significant for 12 out of the 17 subjects. * p < 0.05, ** p < 0.01.*

(2) Based on alpha power: Alpha power is a well-established measure of visual attention. We sought to use alpha power here to test if it would reliably inform whether the visual speech was congruent with the attended audio or congruent with the ignored speech. Using the trial-averaged alpha power across the whole scalp, the attended speech modality was classified using LDA with an accuracy of 95.1 ± 8.2% ($p=2.9x10^{-4}$; Fig. 3.6A). Indeed, the classification accuracy was significant for every subject (Fig. 3.6B). We were then interested to see what scalp channels were driving the classification performance. To do this we re-ran the LDA for each channel using clusters of the eight nearest channels. This revealed that channels in parieto-occipital regions underlie the

ability of the classifier to distinguish between attended AVi and attended AVc trials (Fig. 3.6C). In addition, we calculated the difference in raw alpha power across the two conditions and found that the parieto-occipital alpha power is higher when attending AVi compared with attending AVc speech (grand average, Fig. 3.6D).

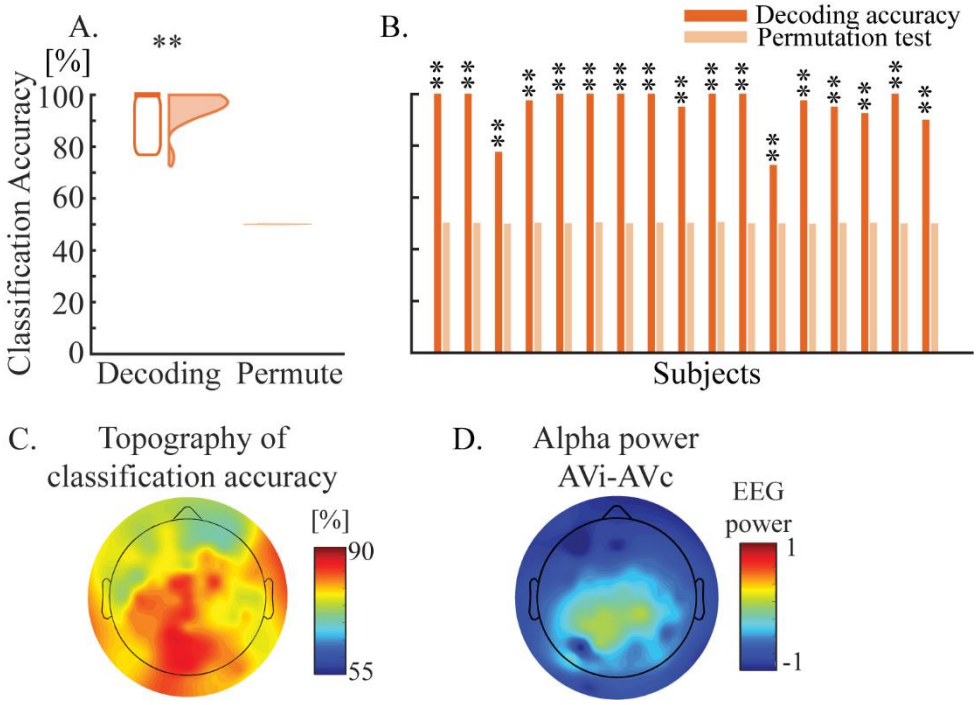


*Figure 3.6 Classification of attended modality using single trial alpha power. A. Classification accuracy is significantly greater than chance which was determined using a 1,000 repetitions permutation test. B. Single subject classification scores are significant for all of the seventeen subjects. C. Classification accuracy using eight channel clusters show that the parieto-occipital region is the most informative regarding the attended speech modality. D. Grand average alpha power difference for AVi and AVc conditions. **p < 0.01.*

Eye movement analysis. To ensure that the differences in decoding performance based on the relevance of visual information are not generated by differences in gaze patterns between the two conditions we compared the time spent looking at the speakers eyes ROI vs the mouth ROI across conditions for each subject (as in Gurler *et al.*, 2015). We found that in both conditions subjects looked at the eyes (as instructed) more than the mouth of the speaker (AVi: percent time viewing eyes: 40.9 ± 23%, percent time viewing mouth: 5.8 ± 10.2%; $p=6.1x10^{-4}$; AVc: percent time viewing eyes: 30.4 ± 15%, percent time viewing mouth: 11.9 ± 11.2%; p=0.02). Nonetheless, subjects did tend to look more at the mouth during the AVc condition compared with AVi (p=0.035). To test if this increase in mouth gazing during the AVc condition benefitted the accuracy of our stimulus reconstruction, we calculated the correlation (Pearson's r) between the amount of time spent looking at the mouth and the reconstruction accuracy. We found no correlation

between the reconstruction accuracies and the mouth viewing durations (p=0.98).

3.4. Discussion

In recent years, methods for decoding attention to natural speech have been heavily investigated (O'Sullivan *et al.*, 2014; Mirkovic *et al.*, 2015; Akram *et al.*, 2016; Fuglsang *et al.*, 2017; O'Sullivan *et al.*, 2017; Denk *et al.*, 2018; Miran *et al.*, 2018). Studies examining the influence of visual speech and its effect on both attending and ignoring the corresponding audio speech however are lacking. Taking a naturalistic listening scenario, we show that modality-specific decoders based on reconstruction accuracy are necessary for optimal decoding of selective attention. Furthermore, we find that alpha power is a reliable measure for determining the relevance of the visual speech.

Attending the speaker's face enhances speech comprehension and neural tracking of the acoustic envelope. Our finding that speech comprehension is significantly better for the case of attending AVc speech than for attending AVi speech can be partly explained by evidence that congruent visual speech improves intelligibility in noisy environments (Sumby & Pollack, 1954; Grant & Seitz, 2000; Munhall *et al.*, 2004). In addition, it has been shown that spatially and temporally congruent sensory inputs are more salient and can capture an individual's attention more readily than sensory inputs that are incongruent (Driver, 1996; Van der Burg *et al.*, 2008; 2009). This suggests that AVc speech may capture attention in a bottom-up manner, making it more difficult for subjects to sustain attention to the AVi speaker.

These findings also correspond with the neural activity where we find an enhancement of AVc speech tracking compared with neural tracking of AVi speech. Specifically, we find that reconstruction accuracy of the acoustic envelope for AVc speech is greater than for AVi speech. This fits well with findings that visual speech enhances auditory speech tracking (Schroeder *et al.*, 2008; Crosse *et al.*, 2015a; Peelle & Sommers, 2015), as well as evidence for a reduction in envelope tracking when the speech is incongruent (Crosse *et al.*, 2015a). One of the reasons that congruent visual speech benefits the cortical tracking of the envelope is that visual speech predominantly leads audio speech (Chandrasekaran *et al.*, 2009; Schwartz & Savariaux, 2014), and so provides relevant cues to the upcoming audio speech which you are trying to attend. As well as that, there is evidence to suggest that visual cortex responds to visual speech in a way that is correlated with the acoustic envelope of that speech – and may engage in higher-level speech processing (Schepers *et al.*, 2015; O'Sullivan *et al.*, 2017; Hauswald *et al.*, 2018).

Although the AVc decoder weights in figure 3E do not show strong weightings over occipital scalp, the single subject decoder weights revealed strong positive occipital weights for twelve subjects and negative weights for five subjects (Appendix A). Thus, it appears that occipital activity does contribute to the envelope reconstructions but the nature of this contribution seems to differ among subjects. On top of the contribution from visual regions, there are additional brain regions which have been shown to be actively involved in integrating AVc speech including STS, SMG, IPS and pre-frontal cortex (Calvert *et al.*, 2000; Miller & D'Esposito, 2005; Bernstein *et al.*, 2008; Ozker *et al.*, 2018) which may also be captured by the AVc decoder. The limited spatial resolution of EEG however restricts us to only speculating on the underlying brain regions which drive the scalp recorded signals.

Stimulus reconstruction decodes attention to audiovisual speech and eavesdropping. In examining stimulus reconstruction accuracies we replicate findings that low-frequency EEG tracks the attended envelope to a much greater extent than the unattended envelope (Ding & Simon, 2012a; Ding & Simon, 2012b; Golumbic *et al.*, 2013a; O'Sullivan *et al.*, 2014). Specifically, we show that for attending AVc speech as well as for eavesdropping conditions, the stimulus reconstruction is more correlated with the attended than the ignored speech envelope. The ability to decode attention to AVc speech was unsurprising given previous success of decoding attention to audio-only speech (O'Sullivan *et al.*, 2014), however, it was unclear that decoding attention while eavesdropping would be successful.

Indeed, the fact that we can successfully decode attention to AVi speech may be somewhat explained by the use of top-down attention to suppress the impact of the incongruent visual speech on auditory processing of the attended stream (for e.g., see Alsius *et al.*, 2005; Morís Fernández *et al.*, 2015). As well as this, selective attention must sufficiently enhance tracking of the AVi speech in order to overcome the fact that neural activations relating to the unattended congruent audio and visual speech are present (Talsma & Woldorff, 2005; Talsma *et al.*, 2010).

Modality-specific decoders are necessary for optimal decoding of selective attention to natural speech. At the trial average level, both decoders generate an envelope reconstruction that is correlated with the attended envelope significantly more than the unattended envelope – regardless of the attended speech-modality. However, at the single trial level we find that cross-modal decoding performs poorly (Fig. 4A). This suggests that the activation patterns across the scalp used for decoding attention to AVc speech

differ from those for attending AVi speech. In particular, it is the AVc decoding performance that degrades most severely for cross-decoding which may be attributed to the advantage of the AVc decoder in its ability to capture multisensory processing of the attended speech which is not present for the AVi decoder. This result is in line with the wealth of research identifying differences in responses for unimodal versus multi-modal speech (Besle *et al.*, 2004a; McGettigan *et al.*, 2012) as well as evidence for specific brain regions activated for integrating congruent audio and visual speech (Calvert *et al.*, 2000; Miller & D'Esposito, 2005; Bernstein *et al.*, 2008; Nath & Beauchamp, 2011; Schepers *et al.*, 2015; Zhu & Beauchamp, 2017; Ozker *et al.*, 2018).

This highlights the need for an alternative approach to decoding attention to speech in natural environments, when we don't know where the subject is looking. We resolved this issue by performing LDA on the envelope reconstruction correlations produced by both decoders which we found to be successful at determining the attended speech modality (Fig. 3.5A).

Alpha power of EEG can be used to detect attention to the speaker's face. Alpha power has a long history of association with the attention (Pfurtscheller & Klimesch, 1991; for review, see: Klimesch, 2012). In particular, alpha power has been shown to be important for audio (Kerlin *et al.*, 2010; Ahveninen *et al.*, 2013; Wöstmann *et al.*, 2016; Bednar & Lalor, 2018) and visual spatial attention (Foxe *et al.*, 1998; Worden *et al.*, 2000; Fu *et al.*, 2001; Frey *et al.*, 2014; Mazaheri *et al.*, 2014). However, the usefulness of this measure in attention decoding of natural, multi-modal speech has not been shown before. Here, we find that the alpha power over parieto-occipital channels is a robust measure for determining the attended speech modality (Fig. 3.6A-D). We attribute this finding to the active suppression of visual speech processing while attending incongruent audio, indexed by an increase in alpha power over parieto-occipital regions. In contrast, there is a decrease in alpha power over the same regions when attending congruent audiovisual speech, due to being actively engaged in fully exploiting the visual speech to aid attentional selection to that speaker. This finding is in line with evidence that increases in parieto-occipital alpha power suppresses processing of distracting visual information (Kelly *et al.*, 2006) as well as reports that it may index the reliability of visual speech (Shatzer *et al.*, 2018). Another possible explanation for these findings is due to the spatial separation of the two speakers (30°), given findings that alpha power tracks the spatial location of a sound (Feng *et al.*, 2017; Bednar & Lalor, 2018). These studies however, report a lateralization of alpha power across parietal scalp regions in response to a

spatialized sound. Given the symmetrical nature of the alpha power response we see here in parietal scalp, it is unlikely that our ability to decode attention based on alpha power is driven by the application of auditory spatial attention.

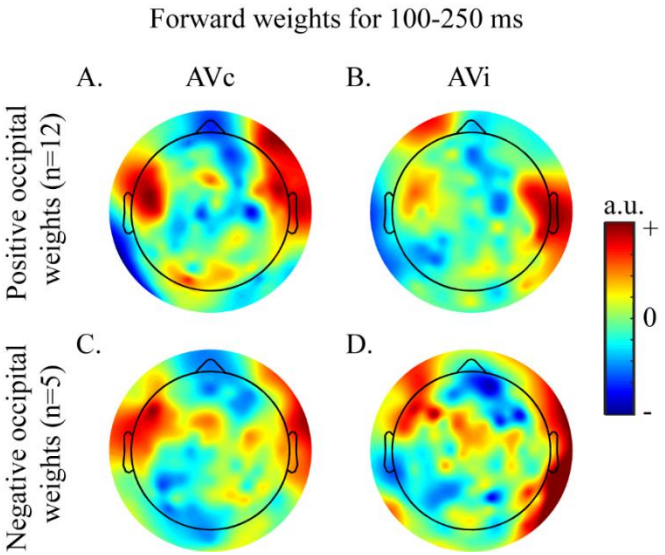
Limitations. It is important to point out some limitations of the current study. The angle of presentation across conditions (0° for AVc speech and 30° for AVi speech) should be counterbalanced to account for any possible ear bias in the decoders (Das *et al.*, 2016). Nonetheless, our cross-decoding effects are quite large (drops by 9.4% and 23%) for a 30° angle of separation between speakers. While Das *et al.* (2016) report a drop of ~11% for the 'other ear' decoder this is likely to be the maximum effect for an ear bias since in that experiment the speakers were presented at ±90°.

3.5. Conclusion

In this study we examined the influence of visual speech on the cortical tracking of the acoustic envelope as a function of the relevance of the visual information. We found that the activation patterns captured on the scalp are different for the case when the visual information is relevant compared to when it is distracting. We also found that alpha power in parieto-occipital regions was modulated by the relevance of the visual speech. This demonstrates that the relevance of visual speech modulates cortical activations when attending to speech in a multisensory cocktail party environment.

One limitation of this study is that we could not separate the cortical contributions of visual speech processing from effects of multisensory integration of the audiovisual speech signal. In the next study, we aim to address this by also recording EEG activity in response to the unisensory stimuli (A-only and V-only). This will allow us to isolate a marker of audiovisual interactions by accounting for the cortical activation caused by the unisensory stimuli. We then plan to test at what cortical stage of speech processing these interactions occur (i.e., at the acoustic vs linguistic level) and how these interactions may differ when there is added noise in the environment.

3.6. Appendix



Appendix A. Decoder weights for AVc and AVi split into two groups based on observation of the single subject topographies. The AVc decoder shows occipital weights in both groups of subjects which are opposite in polarity whereas the AVi decoder shows no weighting of occipital regions for either group. For the AVc decoder one group (n=12) shows positive weights over occipital regions at time lags of 100-250 ms, and the other group (n=5) shows negative weights over the same region for these time lags of interest. Both positive and negative weights contribute to the attended envelope reconstructions but when averaged across all subjects it appears as if there is little to no weighting of occipital regions.

Chapter 4.

Neurophysiological indices of audiovisual speech integration are enhanced at the phonetic level for speech in noise

4.1. Introduction

One prominent theory of speech perception is that speech is processed in a series of computational steps that follow a hierarchal structure, with different cortical regions being specialised for processing different speech features (Scott & Johnsrude, 2003; Hickok & Poeppel, 2007; DeWitt & Rauschecker, 2012). One key question is how visual input influences processing within this hierarchy.

Behavioral studies have shown that seeing the face of a speaker improves speech comprehension (Sumby & Pollack, 1954; Grant & Seitz, 2000; Ross *et al.*, 2007). This behavioral advantage is thought to derive from two concurrent processing modes: a correlated mode, whereby visual speech dynamics provide information on auditory speech dynamics; and a complementary mode, where visual speech provides information on the articulatory patterns generating the auditory speech (Campbell, 2008). It seems plausible that the information provided by these two modes would influence levels of the auditory hierarchy differently. Indeed, this idea aligns well with a growing body of evidence indicating that audiovisual (AV) speech integration likely occurs over multiple stages (Schwartz *et al.*, 2004; van Wassenhove *et al.*, 2005; Eskelund *et al.*, 2011; Baart *et al.*, 2014; Peelle & Sommers, 2015). One recent perspective (Peelle & Sommers, 2015) suggests that these stages could include an early stage, where visual speech provides temporal cues about the acoustic signal (correlated mode), and a later stage, where visual cues that convey place and manner of articulation could be integrated with acoustic information to constrain lexical selection (complementary mode). Such early-stage integration could be mediated by direct projections from visual cortex that dynamically affect the sensitivity of auditory cortex (Calvert *et al.*, 1997; Grant & Seitz, 2000; Tye-Murray *et al.*, 2011; Okada *et al.*, 2013), whereas for later-stage integration, articulatory visual cues could be combined with acoustic information in supramodal regions such as the STS (Beauchamp *et al.*, 2004; Kayser & Logothetis, 2009; Zhu & Beauchamp, 2017; Karas *et al.*, 2019).

While the evidence supporting multiple stages of audiovisual speech integration is compelling, there are several ways in which this multistage model needs to be further developed. First, much of the supporting evidence has been based on experiments involving simple (and often illusory) syllabic stimuli or short segments of speech. This has been very valuable, but it also seems insufficient to fully explore how a correlated mode of audiovisual integration might derive from dynamic visual cues impacting auditory cortical processing. Testing the model with natural speech will be necessary (Theunissen *et al.*, 2000; Hamilton & Huth, 2018). Second, directly indexing neurophysiological representations of different acoustic and articulatory features will be important for validating and further refining the key idea that integration happens at different stages. And third, it will be important to test the hypothesis that this multistage model is flexible, whereby the relative strength of integration effects at different stages might depend on the listening conditions and the availability of visual information.

These are the goals of the present manuscript. In particular, we aim to build on recent work that examined how visual speech affected neural indices of audio speech dynamics using naturalistic stimuli (Luo *et al.*, 2010; Golumbic *et al.*, 2013b; Crosse *et al.*, 2015a; Crosse *et al.*, 2016b). We aim to do so by incorporating ideas from recent research showing that EEG and MEG are sensitive not just to the acoustics of speech, but also to the processing of speech at the level of phonemes (Di Liberto *et al.*, 2015; Khalighinejad *et al.*, 2017; Brodbeck *et al.*, 2018). This will allow us to derive indices of dynamic natural speech processing at different hierarchical levels and to test the idea that audiovisual speech integration occurs at these different levels, in line with the multistage model (Peelle & Sommers, 2015). Finally, we also aim to test the hypothesis that, in the presence of background noise, there will be a relative increase in the strength of AV integration effects in EEG measures of phoneme-level encoding, reflecting an increased reliance on articulatory information when speech is noisy. To do all this, we introduce a new framework for indexing the electrophysiology of audiovisual speech integration based on canonical correlation analysis (CCA).

The findings described in this chapter have been submitted as a research article: O'Sullivan, A., Crosse, M. J., Di Liberto, G. M., de Cheveigné, A., Lalor, E. C. (2020). "Neurophysiological indices of audiovisual speech integration are enhanced at the phonetic level for speech in noise." *Journal of Neuroscience*, in review.

4.2. Material and Methods

The EEG data analyzed here were collected as part of previous studies published by

Crosse et al. (2015a; 2016b).

Participants. Twenty-one native English speakers (eight females; age range: 19-37 years) participated in the speech in quiet experiment. Twenty-one different participants (six females; age range: 21-35) took part in the speech in noise experiment. Written informed consent was obtained from each participant beforehand. All participants were native English speakers, were free of neurological diseases, had self-reported normal hearing, and had normal or corrected-to-normal vision. The experiment was approved by the Ethics Committee of the Health Sciences Faculty at Trinity College Dublin, Ireland.

Stimuli and procedure. The speech stimuli were drawn from a collection of videos featuring a trained male speaker. The videos consisted of the speaker's head, shoulders, and chest, centered in the frame. The speech was conversational-like and continuous, with no prolonged pauses between sentences. Fifteen 60 s videos were rendered into 1280x720 pixel movies in VideoPad Video Editor (NCH Software). Each video had a frame rate of 30 frames per second, and the soundtracks were sampled at 48 kHz with 16-bit resolution. The intensity of each soundtrack, measured by root mean square, was normalized in MATLAB (MathWorks). For the speech in noise experiment, the soundtracks were additionally mixed with spectrally matched stationary noise to ensure consistent masking across stimuli (Ding *et al.*, 2013; Ding & Simon, 2013) with SNR of -9 dB. The noise stimuli were generated in MATLAB using a 50th-order forward linear predictive model estimated from the original speech recording. Prediction order was calculated based on the sampling rate of the soundtracks (Parsons, 1987). The stimuli used in this study will be made available upon reasonable request to the author or advisor.

In both experiments, stimulus presentation and data recording took place in a dark sound attenuated room with participants seated at a distance of 70 cm from the visual display. Visual stimuli were presented on a 19 inch CRT monitor operating at a refresh rate of 60 Hz. The face of the speaker subtended a visual angle of 12°. Audio stimuli were presented diotically through Sennheiser HD650 headphones at a comfortable level of ~65 dB. Stimulus presentation was controlled using Presentation software (Neurobehavioral Systems). For the speech in quiet experiment each of the 15 speech passages was presented seven times, each time as part of a different experimental condition. Presentation order was randomized across conditions, within participants. While the original experiment had seven conditions, here we focus only on three conditions audio-only (A), visual-only (V) and congruent audio-visual (AVc). For the speech in noise experiment, however, there were only 3 conditions (A, V and AV) and so the passages

were ordered 1-15 and presented 3 times with the condition from trial-to-trial randomized. This was to ensure that each speech passage could not be repeated in another modality within 15 trials of the preceding one. Participants were instructed to fixate on either the speaker's mouth (V, AVc) or a gray crosshair (A) and to minimize eye blinking and all other motor activity during recording.

For both experiments participants were required to respond to target words via button press. Before each trial, a target word was displayed on the monitor until the participant was ready to begin. All target words were detectable in the auditory modality except during the V condition, where they were only visually detectable. A target word was deemed to have been correctly detected if subjects responded by button press within 0–2 seconds after target word onset. In addition to detecting target words, participants in the speech-in-noise experiment were required to rate subjectively the intelligibility of the speech stimuli at the end of each 60-s trial. Intelligibility was rated as a percentage of the total words understood using a 10-point scale (0–10%, 10–20%, ... 90–100%).

EEG acquisition and preprocessing. The EEG data were recorded using an ActiveTwo system (BioSemi) from 128 scalp electrodes and two mastoid electrodes. The data were low-pass filtered on-line below 134 Hz and digitized at a rate of 512 Hz. Triggers indicating the start of each trial were recorded along with the EEG. Subsequent preprocessing was conducted off-line in MATLAB; the data were detrended by subtracting a 50th-order polynomial fit using a robust detrending routine (de Cheveigné & Arzounian, 2018). The data were then bandpass filtered using second-order, zero phase-shift Butterworth filters between 0.3-30 Hz, downsampled to 64 Hz, and rereferenced to the average of the mastoid channels. Channels contaminated by noise were recalculated by spline-interpolating the surrounding clean channels in EEGLAB (Delorme & Makeig, 2004).

Indexing neurophysiological speech processing at different hierarchical levels. Because our aim was to examine how visual information affects the neural processing of auditory speech at different hierarchical levels, we need to derive separable EEG indices of processing at these levels. To do this, we followed work from Di Liberto et al. (2015) who modeled EEG responses to speech in terms of different representations of that speech. Specifically, they showed that EEG responses to speech were better predicted using a representation of speech that combined both its low-level acoustics (i.e., its spectrogram) and a categorical representation of its phonetic features. The underlying idea is that EEG responses might reflect the activity of neuronal populations in auditory

cortex that are sensitive to spectrotemporal acoustic fluctuations and of neuronal populations in association cortices (e.g., the superior temporal gyrus) that may be invariant to spectrotemporal differences between utterances of the same phoneme and, instead, are sensitive to that phoneme category itself. As such, for the present study, we calculated two different representations of the acoustic speech signal.

1. **Spectrogram:** This was obtained by first filtering the speech stimulus into 16 frequency bands between 80 and 3000 Hz using a compressive gammachirp auditory filter bank that models the auditory periphery. The gammachirp toolbox was obtained by direct request to the corresponding author on the paper (Irino & Patterson, 2006). Then the amplitude envelope for each frequency band was calculated using the Hilbert transform, resulting in 16 narrow band envelopes forming the spectrogram representation.
2. **Phonetic features:** This representation was computed using the Prosodylab-Aligner (Gorman *et al.*, 2011) which, given a speech file and the corresponding textual orthographical transcription, automatically partitions each word into phonemes from the American English International Phonetic Alphabet (IPA) and performs forced-alignment (Yuan & Liberman, 2008), returning the starting and ending time-points for each phoneme. Manual checking of the alignment was then carried out and any errors corrected. This information was then converted into a multivariate time-series that formed a binary array, where there is a one representing the onset and duration of each phoneme and zeros everywhere else. To describe the articulatory and acoustic properties of each phoneme a 19-dimensional phonetic feature representation was formed using the mapping defined by (Chomsky & Halle, 1968; Mesgarani *et al.*, 2014). This involves mapping each phoneme (e.g., /b/) into a set of phonetic features (e.g., bilabial, plosive, voiced, obstruent) and results in a phonetic feature matrix of ones and zeros that is of dimension 19 (which is the number of phonetic features) by time.

Canonical correlation analysis. We wished to see how these different speech representations might be reflected in EEG activity. Previous related research has relied on a regression-based approach that aims to reconstruct an estimate of some univariate feature of the speech stimulus (e.g., its amplitude envelope) from multivariate EEG responses (Crosse *et al.*, 2015a; Crosse *et al.*, 2016b). However, because we have multivariate speech representations, we sought to use a method based on canonical correlation analysis (CCA, Hotelling, 1936; de Cheveigne *et al.*, 2018) which was

implemented using the NoiseTools toolbox (http://audition.ens.fr/adc/NoiseTools/).

CCA works by rotating two given sets of multidimensional data into a common space in which they are maximally correlated. This linear transformation is based on finding a set of basis vectors for each of the given data sets such that the correlation between the variables, when they are projected on these basis vectors, is mutually maximized. In our case, our two data sets are the multidimensional stimulus representation, $X(t)$, of size $T \times J_1$, where T is time and J_1 is the number of features in that representation ($J_1 = 16$ frequency bands of a spectrogram, or $J_1 = 19$ phonetic features), and an EEG data matrix $Y(t)$ of size $T \times J_2$, where T is time and $J_2 = n\,x\,\tau$, where n is the number of EEG channels (128) and τ is the number of time-lags. The reason for using multiple time lags is to allow for the fact that a change in the stimulus impacts the EEG at several subsequent time lags. In our analysis we included time-lags from 0-500 ms, which at a sampling rate of 64 Hz resulted in 32 time-lags. For these two data matrices, CCA produces transform matrices A and B of sizes $J_1 \times J_0$ and $J_2 \times J_0$ respectively, where J_0 is at most equal to the smaller of J_1 and J_2. The optimization problem for CCA is formulated as a generalized eigenproblem with the objective function:

$$\begin{pmatrix} 0 & C_{XY} \\ C_{YX} & 0 \end{pmatrix} \begin{pmatrix} A \\ B \end{pmatrix} = \rho^2 \begin{pmatrix} C_{XX} & 0 \\ 0 & C_{YY} \end{pmatrix}$$

Where C_{XY} is the covariance of the two datasets X and Y and C_{XX} and C_{YY} are the autocovariances, and ρ is the components correlation. Ridge regularization can be performed on the neural data to prevent overfitting in CCA as follows (Vinod, 1976; Leurgans *et al.*, 1993; Cruz-Cano & Lee, 2014; Bilenko & Gallant, 2016):

$$\begin{pmatrix} 0 & C_{XY} \\ C_{YX} & 0 \end{pmatrix} \begin{pmatrix} A \\ B \end{pmatrix} = \rho^2 \begin{pmatrix} C_{XX} & 0 \\ 0 & C_{YY} + \lambda I \end{pmatrix}$$

Only the EEG by time-lags matrix (C_{YY}) is regularized since the stimulus representations are of low-dimensionality (16-19 dimensions). The rotation matrices (A and B) are learned on all trials except one and are then applied to the left-out data which produces canonical components (CC's) for both the stimulus representation and the EEG using the following equation:

$$X_{J1}(t)\mathrm{A} \rightarrow CC_{stim} \rightarrow \rho \leftarrow CC_{resp} \leftarrow Y_{J2}(t)B$$

The rotation weights A and B are trained to find what stimulus features influence the EEG and what aspects of the EEG are responsive to the stimulus, respectively, in order to maximize the correlation between the two multivariate signals. When the rotation weights

are applied to the left-out data we get the canonical components of the stimulus (CC_{stim}) and of the response data (CC_{resp}). The first pair of canonical components define the linear combinations of each data set with the highest possible correlation. The next pair of CCs are the most highly correlated combinations orthogonal to the first, and so-on (de Cheveigne *et al.*, 2018).

Indexing multisensory integration using CCA. We wished to use CCA to identify any neural indices of multisensory integration during the AV condition beyond what might be expected from the unisensory processing of audio and visual speech. We sought to do this by modelling the encoding of the speech representations in the audio-only (A) and visual-only (V) EEG data and then investigating if there is some difference in the speech-related activity in the AV EEG data which is not present in either of the unisensory conditions. In other words, and in line with a long history of multisensory research (Berman, 1961; Stein & Meredith, 1993; Klucharev *et al.*, 2003; van Wassenhove *et al.*, 2005; Besle *et al.*, 2008), we sought to compare AV EEG responses to A+V EEG responses using CCA and to attribute any difference (i.e., AV – (A+V)) to multisensory processing.

To implement this, we summed the EEG data from matching audio-only and visual-only stimuli (i.e., audio-only and visual-only stimuli that came from the same original AV video; Fig. 4.1A). Thus, for each of the original 15 videos, we ended up with AV EEG responses and corresponding A+V EEG responses. Then, we used CCA to relate the multivariate speech representations (spectrogram + phonetic features) to each of these two EEG responses (AV and A+V). This provides two sets of rotation matrices, one between the stimulus and the AV EEG and one between the stimulus and the A+V EEG.

Now, if the scalp recorded EEG activity for the AV condition is simply the auditory and visual modalities being processed separately with no integration occurring, then the A+V and AV EEG responses should be essentially identical. And we would then expect the rotation matrices learned on the A+V EEG data to be identical to those learned on the AV EEG data. Carrying this logic even further, we would then expect to see no differences in the canonical correlation values obtained from the AV data when using the CCA rotation matrices found by training on the A+V EEG data compared with the matrices found by training on the AV EEG data (Fig. 4.1B). In other words, we compared the correlation values obtained when we applied the AV weights (i.e., the A and B matrices found by training on AV EEG data) to left-out AV data, with the correlation values obtained when applying the A+V weights (i.e., the A and B matrices found by training on A+V EEG

data) to the AV data (Fig. 4.1B). If there is some difference in the EEG response dynamics for multisensory (AV) compared with the summed unisensory activity (A+V) then we would expect this to have a significant effect on the canonical correlations since the A+V weights would not capture this whereas the AV weights would. To measure the size of this difference we calculate multisensory gain using the following equation:

$$MSI_{Gain} = \frac{\rho_{AV,AV} - \rho_{A+V,AV}}{|\rho_{AV,AV}| + |\rho_{A+V,AV}|}$$

Where ρ is the canonical components correlation, the first subscript represents the rotations used and the second subscript represents the data on which those rotations are applied (Fig. 4.1B). The difference in performance between the models is normalized since the cortical tracking for very noisy speech is typically much weaker than the tracking of clean speech (Ding & Simon, 2013; Crosse *et al.*, 2016b) and cortical tracking correlation values can also vary substantially across subjects due to differences in cortical folding, skull thickness and scalp thickness. Thus, normalizing the difference in correlations ensures that results from all subjects in both conditions are represented such that they can be compared fairly.

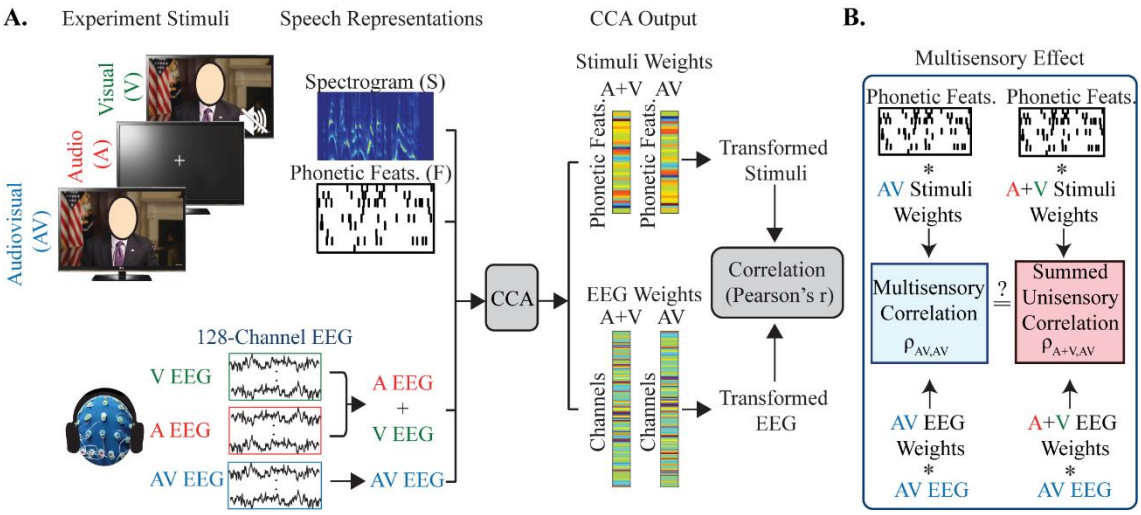


Figure 4.1. Experiment set-up and analysis approach. The face of the speaker is blocked with an oval for publication but was not blocked for the experiment. A. The stimulus representations used are the spectrogram and the phonetic features and are estimated directly from the speech stimuli. Below, the EEG recordings corresponding to each condition. The unisensory A and V EEG are summed to form an A+V EEG data set. The EEG and speech representations are used as inputs to the CCA in order to determine the optimum weights for rotating the EEG and the given stimulus representation for maximizing the correlation between the two. B. The model built using the A+V data is then tested on the left-out AV data in order to determine the presence of a multisensory effect.

Statistical analysis. All statistical comparisons were conducted using non-parametric permutation with 10,000 repetitions such that no assumptions were made about the

sampling distribution (Combrisson & Jerbi, 2015). This was done by randomly assigning the values from the two groups being compared (pairwise for the paired tests, and non-pairwise for the unpaired tests) and calculating the difference between the groups. This process was repeated 10,000 in order to form a null distribution of the group difference. Then the tail of this empirical distribution is used to calculate the p-value for the actual data, and two-tailed tests are used throughout. Where multiple comparisons were carried out p-values were corrected using the False Discovery Rate (FDR) method (Benjamini & Hochberg, 1995). All numerical values are reported as mean ± SD.

4.3. Results

The behavioural results are not shown here but can be found in Crosse et al., 2015a and 2016b for the speech in quiet and speech is noise experiments respectively.

Robust indices of multisensory integration for the speech spectrogram and phonetic features. To investigate the encoding of more complex multivariate representations of the speech stimulus and to isolate measures of multisensory integration at different levels of the speech processing hierarchy we performed CCA on the AV EEG data using the spectrogram and phonetic features, having trained the CCA on (different) AV data and A+V data. We first sought to do this separately for the spectrogram representation and the phonetic representation to see if using either or both of these representations might show evidence of multisensory integration. And we also sought to do this for both our clean speech and noisy speech datasets.

In both conditions (clean and noisy speech) and for both representations (spectrogram and phonetic features) the correlations for the first component were significantly higher than for all other components. This suggests that the first component captures a substantial percentage of the influence of the speech on the EEG data. And, importantly, both representations showed evidence of multisensory integration.

For the spectrogram representation, we found significant multisensory effects (AV>A+V) for the first canonical component for speech in quiet ($p<0.0001$) and the first canonical component for speech in noise ($p<0.0001$, FDR corrected p-values; Fig. 4.2A, B). Indeed we found multisensory effects for 15/16 components for speech in quiet and for 15/16 components for speech in noise.

A similar pattern was observed when examining the stimulus-EEG relationship using the phonetic feature representation of speech. Specifically, we also found multisensory

effects for the first component for clean speech ($p<0.0001$) and for speech in noise ($p<0.0001$, FDR corrected). And we found multisensory effects for 13/18 components for speech in quiet and for all 18 components for speech in noise. Although there are 19 phonetic features, there are only 18 components since CCA cuts off eigenvalues below a threshold, which in this case was 10^{-12} and the last component of phonetic features did not survive this cut-off.

The fact that the spectrogram and phonetic feature representations produced qualitatively similar patterns of multisensory integration was not surprising. This is because both representations are mutually redundant; a particular phoneme will have a characteristic spectrotemporal signature. Indeed, if each utterance of a phoneme were spoken in precisely the same way every time, then the spectrogram and phonetic feature representations would be functionally identical (Di Liberto *et al.*, 2015). But, as we discuss below, in natural speech different utterances of a particular phonemes will have different spectrograms. So, to identify the unique contribution of "higher-level" neurons that are invariant to these spectrotemporal differences and are sensitive to the categorical phonemic features we will need to index the EEG responses that are uniquely explained by the phonetic feature representation whilst controlling for the spectrogram representation. (Please see the section on **Isolating multisensory effects at the spectrotemporal and phonetic levels** below).

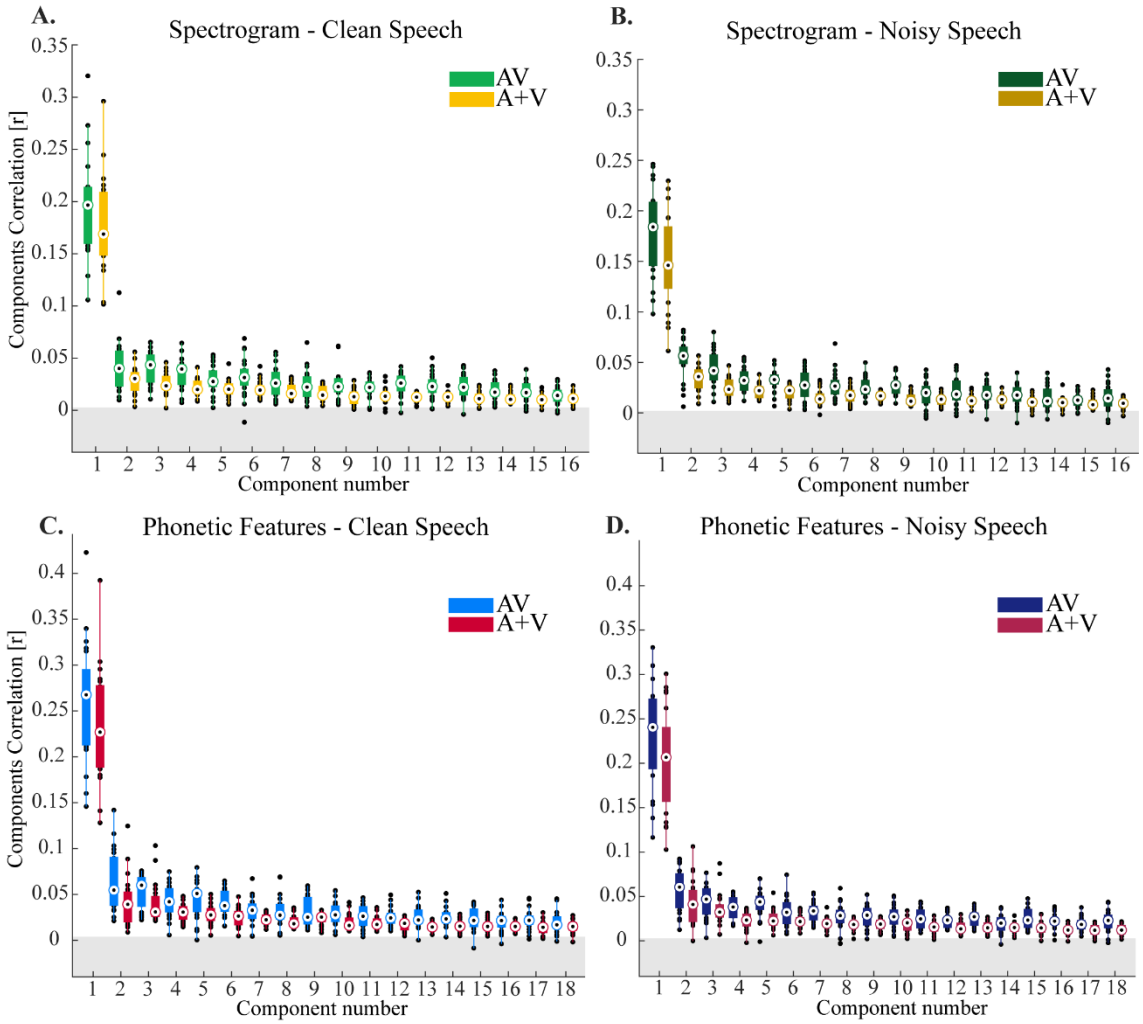


Figure 4.2. CCA analysis using the spectrogram and phonetic feature representation of the speech stimulus. A-B. The canonical correlations for the spectrogram representation for speech in quiet and speech in noise respectively. C-D. canonical correlations for the phonetic feature representation for speech in quiet and speech in noise respectively. The gray band represents an approximate chance level. All AV and A+V components performed above chance level $p<0.0001$ for speech in quiet, and for speech in noise $p<0.0001$. The boxplot function with compact format from MATLAB was used here. The edges of the boxplots display the 25th and 75th percentiles and the median is marked by a black dot inside a white circle. The whiskers extend to the most extreme data points that the algorithm doesn't consider to be outliers and the outliers are plotted individually.

Spatiotemporal analysis of canonical components: increased cross-modal temporal integration and possible increased role for visual cortex for speech in noise. The previous section showed clear evidence of multisensory integration in the component correlation values obtained from CCA. But how can we further investigate these CCA components to better understand the neurophysiological effects underlying these numbers? One way is to examine how these multisensory effects might vary as a function of the time-lag between the stimulus and EEG and how any effects at different time-lags might be represented across the scalp. This is very much analogous to examining the spatiotemporal characteristics of event-related potentials with EEG.

To investigate the spatiotemporal properties of the AV and A+V CCA models we ran the CCA at individual time-lags from -1s to 1.5s. We chose to focus our analysis on the first three canonical components. This was mostly to allow investigation of the dominant first component, but also to check whether or not useful insights might be gleaned from any of the subsequent components. For the first component of the spectrogram model, there was significant differences between AV and A+V at -300-(-200) ms, 0-125 ms and at 500-750 ms for speech in quiet (FDR corrected). For speech in noise there was significant differences at time shifts of -650-(-250) ms, -60-750 ms (Fig. 4.3A, D, FDR corrected). For the second and third components there was no clear pattern to the single time-lag correlation as it was quite flat across all time shifts for both clean and noisy speech (Fig. 4.3B, C, E, and F).

For the phonetic feature representation we found significant differences between AV and A+V at -370-(-125) ms, 90-750 ms for speech in quiet and for speech in noise there was significant differences at time shifts of -300-(-125) ms and 300-750 ms (Fig. 4.3G, J, FDR corrected). For the second component the pattern reflected that of an onset response and while there was no difference between AV and A+V for speech in quiet there was a small window of difference for speech in noise at 0-75ms (Fig. 4.3H, K, FDR corrected). There was no clear pattern to the single time-lag correlation for the third component for both clean and noisy speech (Fig. 4.3I, L). In general, the number of lags at which there was a significant difference between AV and A+V was greater for noisy speech than clean speech, which is consistent with findings in Crosse *et al.* (2016b). The finding of significant differences at negative lags may be due to the fact that the EEG data is time-locked to the onset of the audio and since the visual information often precedes the audio (Chandrasekaran *et al.*, 2009; Schwartz & Savariaux, 2014), the AV data may contain some information about the speech at 'negative' time-lags. Another possibility however, is that the effect at negative lags is due to the autocorrelation of the stimulus and the autocorrelation of the EEG. Nonetheless, in our multi-lag CCA we have only used positive lags (0-500 ms) and so any effects at negative lags will not influence our overall results.

In summary, the single lag analysis reveals multisensory interactions for both stimulus representations at similar ranges of lags – mostly in the 0-500ms range. It also shows the dominance of the first component in capturing the relationship between the EEG and stimulus, however for phonetic features the second component also appears to display a time-locked response. Nonetheless, there are many fewer time-lags for which we see

multisensory interactions in the second and third components. The first component for both spectrogram and phonetic features shows an early peak which is likely related to an onset response at around 100ms post-stimulus. The phonetic features however also shows a second broader peak at around 400ms which is not present for the spectrogram representation.

To visualize the scalp regions underlying these components we calculated the correlation coefficient between each component from the AV models for each time lag with each scalp electrode of the AV EEG. This gives a sense of how strongly the data on each channel has contributed to that component. The spatial pattern for the first component revealed strong contributions from channels over central and temporal scalp for speech in quiet for both the spectrogram and phonetic feature representations. For speech in noise there was an additional correlation with occipital regions, possibly indicating an increased contribution from visual areas to multisensory speech processing in noisy conditions. Occipital channels also made clear contributions for the second and third components for both conditions, however due to the lack of a clear temporal response for these components, we are hesitant to over-interpret this.

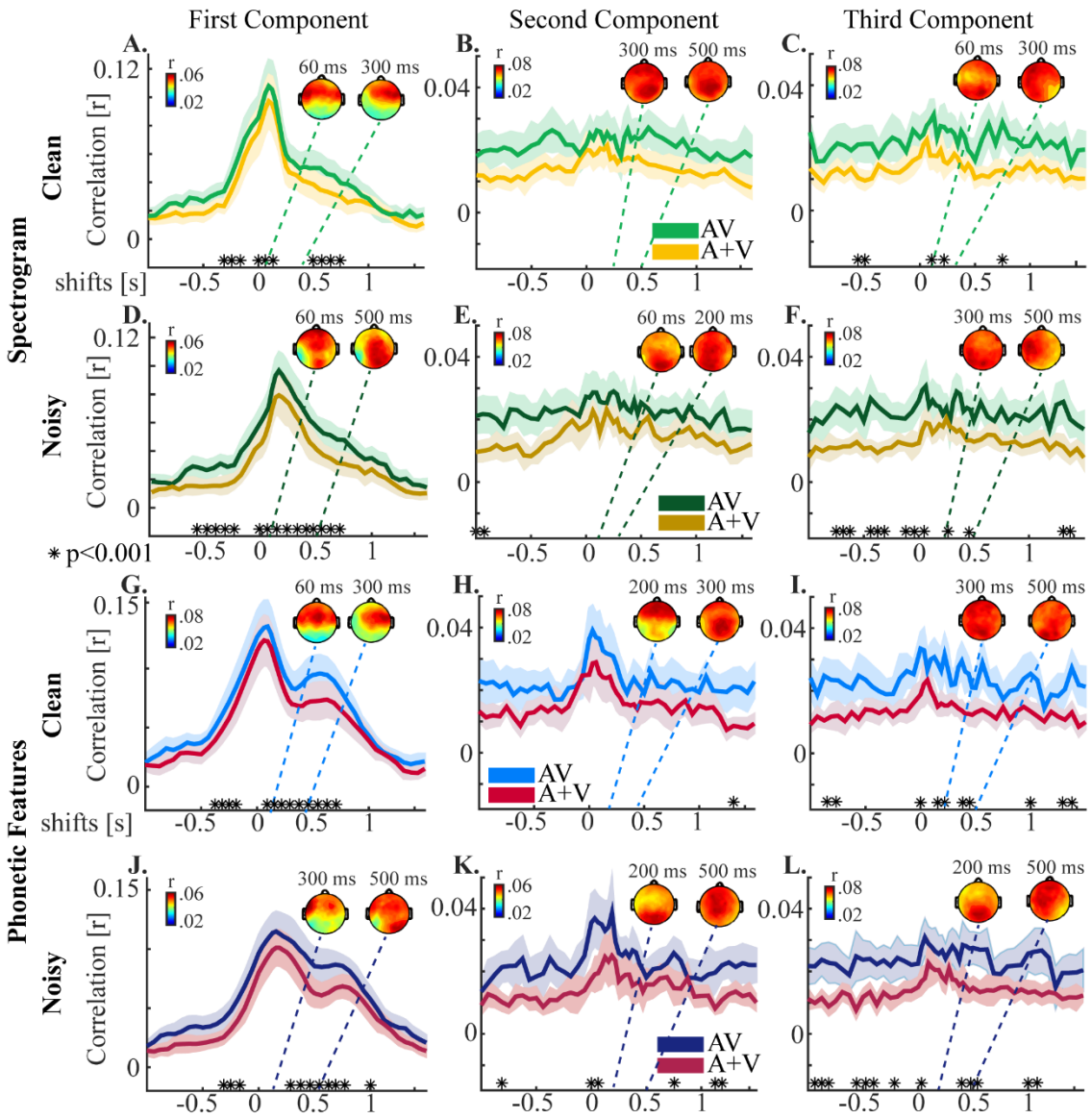


*Figure 4.3. Correlations for the first three canonical components using the spectrogram and phonetic feature representation of the speech stimulus at single time shifts. A-C. The correlations using the spectrogram representation for the first three components using the AV model and the A+V model for speech in quiet, and D-F for speech in noise. G-I The single time shift correlations for the phonetic feature representation between the first three components for the AV and A+V model with the original raw AV EEG data for speech in quiet, and J-I for speech in noise. The respective topographies inset show the corresponding spatial correlations between the corresponding component of the AV model and the AV EEG. *p<0.001 FDR corrected. The shaded region of the line plot marks the 95% confidence interval around the mean.*

Isolating multisensory effects at the spectrotemporal and phonetic levels. As discussed above, the spectrogram and phonetic feature representations are highly correlated with each other. As such, measures of how well each individual representation maps to the EEG (as in Fig. 4.2) are difficult to interpret in terms of multisensory effects at specific hierarchical levels. To pinpoint effects at each specific level, we need to identify the unique contributions of the spectrogram and the phonetic feature

representations to the EEG data. To do this, we first used the forward TRF model (Crosse et al., 2016a) to predict the EEG using one stimulus representation (e.g., the spectrogram). Then we subtracted the predicted EEG from the original EEG signal. Then we fed this residual EEG into the CCA analysis previously described. We performed this analysis for all stimulus representations. To ensure that there were no responses related to the stimulus feature which was regressed out using the TRF which could be extracted by CCA, we re-ran the CCA using the spectrogram and EEG with the spectrogram regressed out. We performed a similar analysis for the phonetic features. In both cases, we found that the correlations are extremely small and close to zero, which so leads us to believe that the partialling out was effective (result not shown).

This should isolate the unique contribution (if any) provided by the phonetic feature representation. Examining such a measure across our 2 conditions (clean speech and noisy speech) allowed us to test the hypothesis that multisensory integration effects should be particularly pronounced at the phonetic feature level for speech in noise. We also performed a similar analysis for the spectrogram representation to test its unique contribution to the multisensory effect in quiet and noise. In this case, we regressed out the phonetic feature representation from the EEG and then related the residual EEG to the spectrogram using CCA. Again, we did this for both speech in quiet and speech in noise, to test for interaction effects on our multisensory integration measures between acoustic and articulatory representations and speech in quiet and noise.

We limited our analysis here to the first canonical component due to its dominant role in the above results, as well as to the fact that it displays a greater consistency across subjects relative to the other components (please see next section of results and Fig. 4.6).

We found that multisensory gain at the level of acoustic processing (unique contribution from spectrogram) was significantly greater than zero for both clean speech and noisy speech (Fig. 4.4A). However, there was no difference in this measure between conditions (Fig. 4.4B; $p=0.75$), suggesting that multisensory integration at the earliest cortical stages was similar for speech in quiet and noise. Meanwhile, multisensory gain at the level of articulatory processing (unique contribution from phonetic features) was also significantly greater than zero for both clean speech and speech in noise (Fig. 4.4C). Importantly however, in line with our original hypothesis, there was a significant difference in this measure between conditions, with MSI gain being larger for speech in noise than speech in quiet (Fig. 4.4D; $p=0.04$). This supports the idea that, when speech is noisy, the impact of complementary visual articulatory information on phonetic feature

encoding is enhanced.

We used R (R Core Team, 2019) and lme4 (Bates *et al.*, 2014) to perform a linear mixed effects analysis (Winter, 2013) with fixed effects of model type (AV vs A+V), stimulus representation (spectrogram or phonetic features) and environment (quiet or noisy) and random effect of subjects. In summary, we found a main effect of model type (driven by larger component correlations for AV vs A+V, $p<0.0001$), and an interaction between stimulus representation and environment (driven by a decrease in correlation values for speech in noise versus speech in quiet for phonetic features, whereas there is no such decrease in correlations for the spectrogram representation across speech conditions, $p<0.0001$), however, the three-way interaction was not significant ($p=0.72$).

We also related the target word detection performance to the multisensory gains for both representations. To do this, we used F1-scores which are calculated as the harmonic mean of precision and recall. This allowed us to investigate if the probability of detecting target words in the multisensory condition exceeded the statistical facilitation produced by the unisensory stimuli (Stevenson *et al.*, 2014). For more information on how this was calculated see Crosse *et al.* (2016b). However, for both representations there was no correlation across subjects between the behavioural gain in target word detection and the multisensory gain calculated from the EEG using CCA (spectrogram: $r=0.004$, $p=0.98$; phonetic features: $r=0.04$, $p=0.86$).

It is difficult to isolate phoneme specific responses from the purely acoustic driven features, given their tightly linked association. To address the possibility that other forms of acoustic representations could explain this result, we re-ran 2 separate analyses using the half-wave rectified spectrogram derivative (Daube *et al.*, 2019) and phonetic onsets (Brodbeck *et al.*, 2018) in place of the phonetic features to test if these representations would lead to similar results as for the phonetic features.

Using the spectrogram derivative we found no difference in the gain between quiet and noisy speech conditions (Fig. 4.4F; $p=0.1$). Similarly, in the case of phonetic onsets we found no effect (Fig. 4.4H; $p=0.4$).

We also related the phonetic features to EEG that had both the spectrogram and spectrogram derivative partialled out. In this case, we found that there was no longer a significant difference in the gain between quiet and noisy speech conditions ($p=0.25$). This is likely due to the fact that the size of the original effect is small due to it coming from differences between two models that are expected to be very similar in the first place, i.e., the AV and A+V models. Therefore regressing out these other representations

reduces the correlation values further, resulting in a reduction in sensitivity to small effects. On top of this we are comparing across different subjects, making our statistics less sensitive than would be the case for a within-subject design. Nevertheless, from fig. 4.4C below it is clear that the phonetic feature representation has noticeably higher correlation values compared with the correlation values for the other representations. This shows that phonetic features are explaining more variance in the EEG. Future work using a within subjects experiment might be more sensitive for testing some of the mechanistic ideas we have proposed in this study.

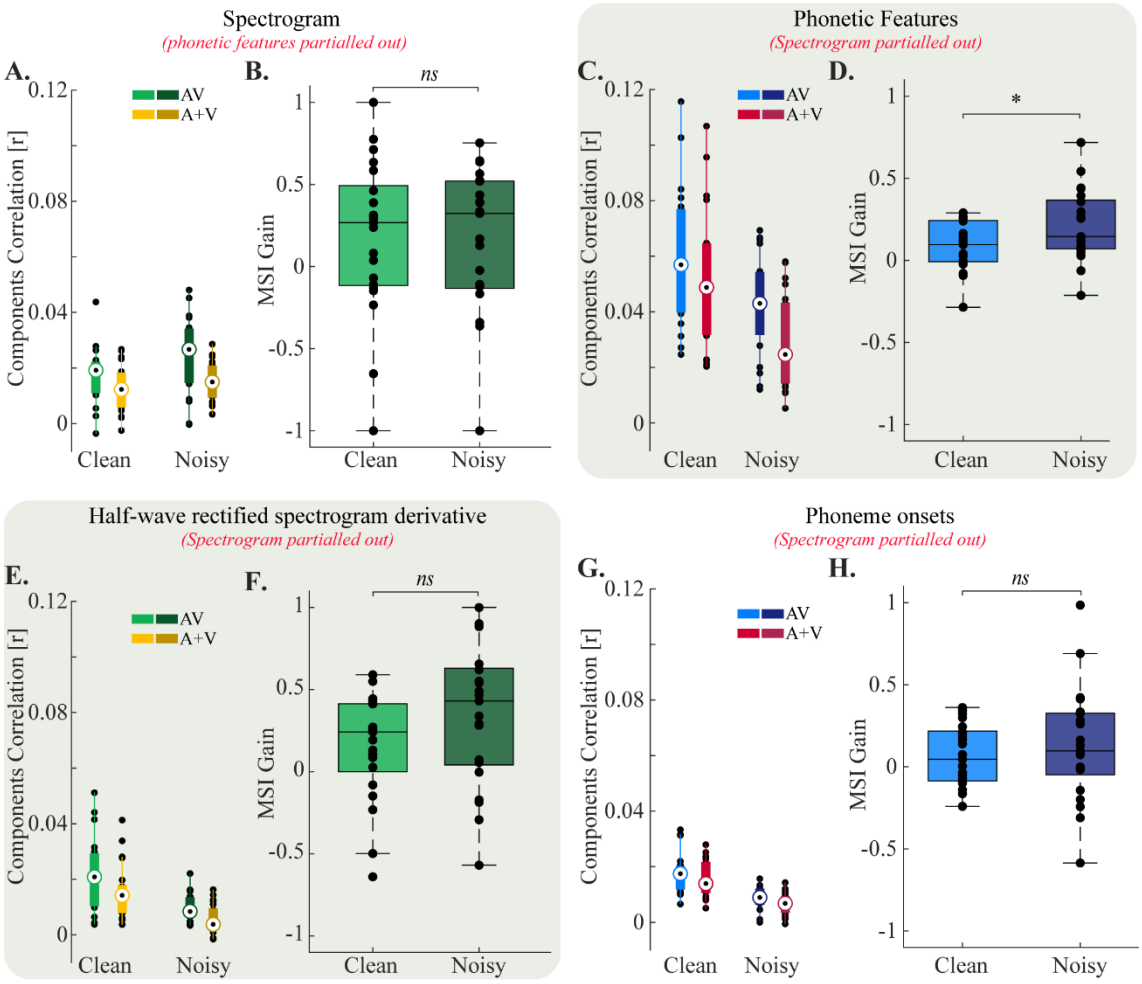


Figure 4.4. Multisensory gain for different speech representations for speech in quiet and in noise. A, B. For the unique spectrogram representation (phonetic features partialled out) there is no difference in gain, p=0.75. C, D. For the unique phonetic features (spectrogram partialled out) we find a difference in gain between conditions, p=0.04. E, F. Using the half-wave rectified spectrogram derivative we found no difference in gain between quiet and noisy conditions (p=0.1) and similarly for phonetic onsets there was no difference across conditions (p=0.4). Two-tailed unpaired permutation tests were used throughout. The boxplot function with compact format from MATLAB was used for figure parts A, C, E, and G. The edges of the boxplots display the 25th and 75th percentiles and the median is marked by a black dot inside a white circle. The whiskers extend to the most extreme data points that the algorithm doesn't consider to be outliers and the outliers are plotted individually. For figure parts B, D, F and H the default boxplot from MATLAB is

used. The only difference in this case is that the median is displayed using a black horizontal line.

We then wanted to examine whether the multisensory integration effects at the phonetic feature level might be driven by specific phonemes. More precisely, we wondered if the effect might be primarily driven by phonemes whose accompanying visual articulations are particularly informative (e.g., /b/, /p/ or /f/ compared to /g/ or /k/). To do this, we tested which phonemes were most correlated with the first canonical component. This involved taking the first component arising from the rotation of the phonetic feature stimulus on the left-out data and then calculating the correlation between this component and the time-series of each phoneme. If there is a high correlation between the component and a particular phoneme then it suggests that this phoneme is strongly represented in that component and it plays a role in driving the stimulus-EEG correlations that we have reported here. This analysis revealed that phonemes such as /p/, /f/, /w/, and /s/ were most strongly represented in the first component. In general, it also showed that consonants were more correlated with the component than vowels (Fig. 4.5A, B), although for speech in noise this effect was slightly less pronounced (Fig. 4.5B). To see these results in terms of visual articulatory features, we grouped the phonemes into visemes (the visual analog of phonemes), based on the mapping defined in (Auer Jr & Bernstein, 1997). This showed that bilabials (/b/,/p/ and /m/) and labio-dentals (/f/, and /v/) were the features most correlated with the first component. Finally, we also checked that these phoneme-component correlations were not simply explainable as a function of the number of occurrences of each phoneme. To check this, we tested for a relationship between the phoneme-component correlations and the number of occurrences of each phoneme (Fig. 4.5B grey line). No correlation between the two was found for either speech in quiet ($p = 0.99$) or speech in noise ($p = 0.71$).

This analysis highlights how different visual-phonetic features contribute to our multisensory effects for phonetic features. In particular we find that bilabials and labiodentals have the highest correlation with the first canonical component suggesting that these features contribute most to the effects shown here. This is line with early work examining the saliency of visual phonetic features which found that place of articulation (i.e., the ability to distinguish between labials, e.g., /p/, /b/ and non-labials e.g., /d/, /t/) was the most salient visual phonetic feature, followed by manner of articulation (i.e., distinguishing between stops, e.g., /d/ and fricatives e.g., /f/) and voicing (Walden *et al.*, 1977).

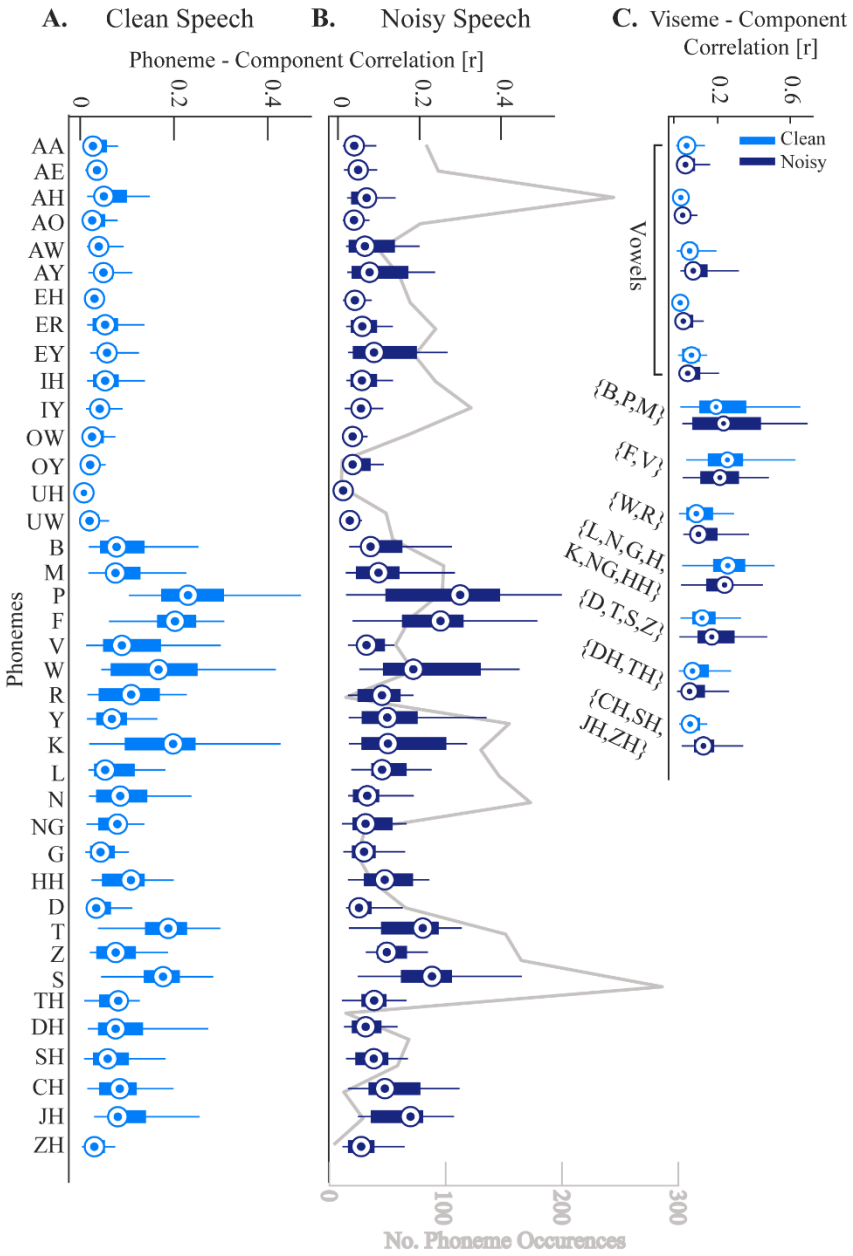


Figure 4.5. Correlation between phonemes and visemes time-series with the first canonical component. A. The correlations for each phoneme from the clean speech data. B. The correlations for each phoneme from the speech in noise data. The grey line plots the number of phoneme occurrences to show that it is not the case that the most frequent phonemes dominate the data. C. The correlation between each viseme (groups of visually similar phonemes) and the first component. The compact version of MATLAB's boxplot function is used here (for description see figure 4 caption).

Consistency of canonical components across subjects. CCA finds stimulus-EEG matrix rotations on a single subject basis. As such, for us to make general conclusions about results gleaned from individual canonical components, we must examine how similar the individual components are across subjects. To do this, we took the components for each subject and calculated the correlation (using Pearson's r) for every subject pair (Fig. 4.6). Given its dominant role in capturing EEG responses to speech, and in the results we have presented above, we were particularly interested in consistency of

component one across subjects.

For the spectrogram, the first components of the AV and A+V models were significantly more correlated across subjects than all other components for clean speech ($p<0.0001$ for both). For speech in noise the first component of the AV model was not significantly more correlated across subjects than the second component ($p=0.055$) but it had a significantly higher correlation than the remainder of the components ($p<0.0001$). The first A+V component for speech in noise was significantly more correlated across subjects than all others ($p<0.001$).

For the phonetic features model, a similar pattern emerged for the clean speech condition with the first components of the AV and A+V models being significantly more correlated across subjects than all other components ($p<0.0001$ for both). For noisy speech, the first component of the AV model was again significantly better than all others ($p<0.02$) and similarly the first component of the A+V model had a higher correlation than all others ($p<0.0001$).

Altogether, we found that the first component is most correlated across subjects, while later components are less correlated. This suggests that the first component, which dominated our results above, is also the most consistent component across subjects and, thus, that it is capturing processes that are general across those subjects. In contrast, the later components, as well as being smaller, are more variable across subjects, and, accordingly, may not be capturing similar underlying processes. This in turn can make results from these later components more difficult to interpret.

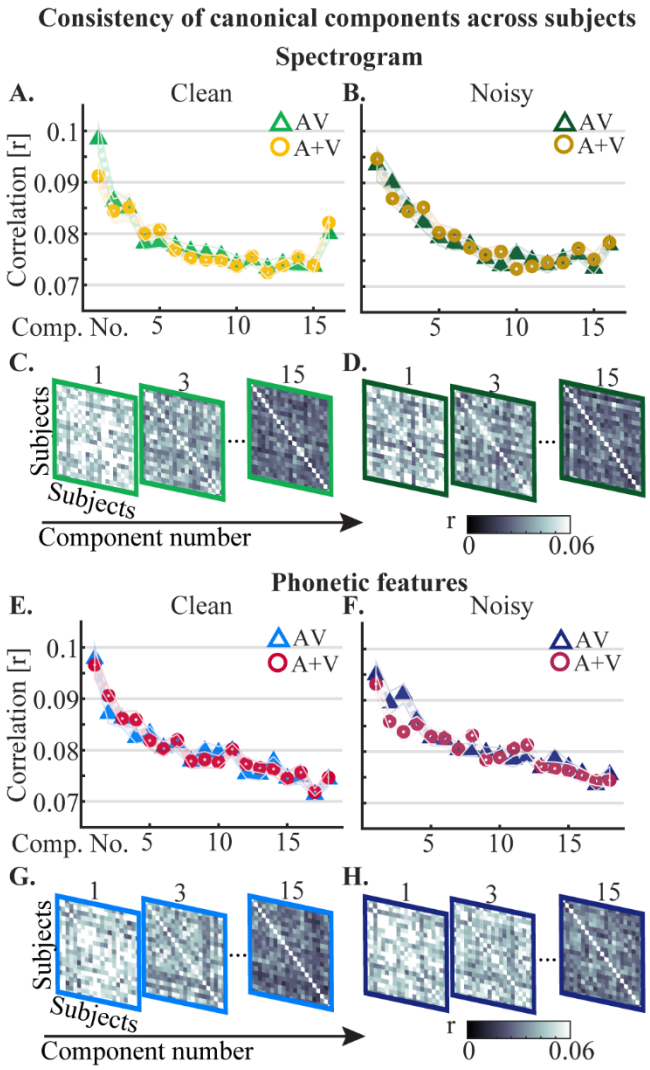


Figure 4.6. Consistency of canonical components across subjects for the spectrogram and the phonetic features after partialling out the other representation. A, B. Correlations of AV and A+V canonical components across subjects for the spectrogram representation for speech in quiet and speech in noise respectively. C, D. Correlation matrices visualizing the reduction in consistency in the AV component activity across subjects as the component number increases for the spectrogram. E, F. The correlation of the canonical components for the phonetic feature representation across subjects, and G, H. the corresponding correlation matrices for speech in quiet and speech in noise respectively.

4.4. Discussion

In this work, we have used a CCA-based framework for relating multivariate stimulus representations to multivariate neural data in order to study the neurophysiological encoding of multidimensional acoustic and linguistic features of speech. Our results show significant audiovisual integration effects on the encoding of the spectrogram and phonetic features of both clean and noisy speech. Importantly, these multisensory effects are enhanced at the phonetic level for speech in noise, supporting the hypothesis that listeners increasingly rely on visual articulatory information when speech is noisy.

Enhanced multisensory integration effects at the phonetic-level of processing for speech in noise. It is well known that the enhancement of auditory speech processing provided by visual speech varies with listening conditions (Ross *et al.*, 2007). However, the details of how visual speech impacts auditory speech processing at different hierarchical levels remains to be fully elucidated. There is a growing body of evidence indicating that AV speech integration likely occurs over multiple stages (Peelle & Sommers, 2015). In particular, it is thought that visual speech provides temporal information about the acoustic speech which can affect the sensitivity of auditory cortex (Grant & Seitz, 2000; Okada *et al.*, 2013), as well as provide complementary cues that contain articulatory information which may be integrated with acoustic information in superior temporal sulcus (STS) (Beauchamp *et al.*, 2004; Kayser & Logothetis, 2009; Nath & Beauchamp, 2011).

In this study, we found that audiovisual speech integration operates differently under different listening conditions (clean vs noisy (-9dB) speech). Specifically, for the encoding of low-level spectrogram features, we found that the integration effects are substantial for both speech in quiet and speech in noise. These integration effects are likely to be primarily driven by modulations of responses in early auditory cortex by temporal information provided by visual speech, which often precedes the auditory speech (Chandrasekaran *et al.*, 2009; Schwartz & Savariaux, 2014). This result is also in line with recent work demonstrating multisensory benefits at the spectrotemporal level elicited by a visual stimulus that did not contain articulatory detail - dissociating the effect from access to higher-level articulatory details (Plass *et al.*, 2019). Furthermore, the lack of any difference in the magnitude of these integration effects between clean and noisy speech conditions suggests that the benefits of visual speech provided at a low-level of processing might be similar regardless of acoustic conditions.

In contrast with this, using a higher-level phonetic feature representation, we found that the AV integration effects are different depending on the acoustic conditions (after regressing out the contribution of the spectrogram). Specifically, we found significantly larger integration effects for phonetic feature encoding in noisy speech than in clean speech. We suggest that this benefit is likely to be driven by an increased reliance on the visual articulations which help the listener to understand the noisy speech content by constraining phoneme identity (Karas *et al.*, 2019). In line with this, we also show that the phonemes that most contribute to these results are those that have particularly informative visual articulations (Fig. 4.5).

While recent research has challenged the notion that scalp recorded responses to speech reflect processing at the level of phonemes (Daube *et al.*, 2019), our findings reveal a sharp dissociation in AV integration effects on isolated measures of acoustic and phonetic processing across listening conditions. This seems difficult to explain based on considering acoustic features alone and seems consistent with the idea of visual articulations influencing the categorization of phonemes (Holt & Lotto, 2010). More generally, we take this as a further contribution to a growing body of evidence for phonological representations in cortical recordings to naturalistic speech (Di Liberto *et al.*, 2015; Khalighinejad *et al.*, 2017; Brodbeck *et al.*, 2018; Yi *et al.*, 2019; Gwilliams *et al.*, 2020).

One brain region likely involved in exploiting the articulatory information when the speech signal is noisy is superior temporal sulcus (STS) which has been shown to have increased connectivity with visual cortex in noisy compared with quiet acoustic conditions (Nath & Beauchamp, 2011). While it remains an open question as to how much speech-specific processing is performed by visual cortex (Bernstein & Liebenthal, 2014), there is a some early evidence supporting the notion that visual cortex might processes speech at the level of categorical linguistic (i.e., phonological) units (O'Sullivan *et al.*, 2017; Hauswald *et al.*, 2018). If true, visual cortex would be in a position to relay such categorical, linguistic information to directly constrain phoneme identity, again, possibly in STS. On top of this it has been shown that frontal cortex selectively enhances processing of the lips during silent speech compared with when the auditory speech is present, suggesting an important role for visual cortex in extracting articulatory information from visual speech cues (Ozker *et al.*, 2018). Thus it is plausible that the greater multisensory gain seen here for phonetic features when the speech is noisy is underpinned by an enhancement of mouth processing in visual cortex which feeds information about the articulations to STS where they influence the online processing of the acoustic speech.

Investigating hierarchical stages of speech processing - CCA captures relationships between multi-dimensional stimuli and EEG. The event related potential (ERP) technique has for a long time been used to advance our understanding of the multisensory integration of speech (Meredith & Stein, 1993; Molholm *et al.*, 2002; Klucharev *et al.*, 2003; van Wassenhove *et al.*, 2005; Saint-Amour *et al.*, 2007; Bernstein *et al.*, 2008; Shahin *et al.*, 2018). However, this approach is ill suited for use with natural, continuous speech stimuli.

More recently, researchers have begun to use methods such as multivariate regression (Crosse *et al.*, 2016a) in the forward direction (predicting neural data from the stimulus, Lalor & Foxe, 2010; Ding & Simon, 2012b; Golumbic *et al.*, 2013b; Di Liberto *et al.*, 2015; O'Sullivan *et al.*, 2017; Broderick *et al.*, 2018) and backward direction (stimulus reconstruction from neural data, Mesgarani *et al.*, 2009; Crosse *et al.*, 2015a; Crosse *et al.*, 2015b; Crosse *et al.*, 2016b), which allows characterization of neural responses to natural speech. However, regression models in their general form allow only univariate-multivariate comparison, whereas with CCA one can relate multivariate stimulus representations (discrete/continuous) to multivariate neural responses. This is a useful advance over current techniques to study speech processing since CCA can use all features (of the stimulus and the neural response data) simultaneously to maximize the correlation between the speech representation and the neural data (de Cheveigne *et al.*, 2018). Importantly, this approach has allowed us to answer questions which we could not do with previous methods, such as the impact of visual speech on auditory speech processing at different stages.

Our results show significant multisensory interaction effects in EEG responses based on the spectrogram and phonetic feature representations of the speech signal and so provides support for the multistage framework for audiovisual speech integration. Examining the relationship between the stimulus representations and EEG data at individual time-shifts reveals a peak in the correlation at around 100 ms post-stimulus for both the spectrogram and phonetic feature representations. This is likely attributable to a sound onset response. For the phonetic feature representation however, there is also a second broad peak at around 300-600 ms whereas for the spectrogram there is no noticeable second peak. There also appears to be a possible delay in the single-lag correlation for speech in noise compared with speech in quiet. While it has previously been shown that responses to noisy speech are delayed relative to those from clean speech (Ding & Simon, 2013), we did not carry out direct comparisons between the latency of the peaks for speech in quiet vs speech in noise here. However, given previous work (Ding & Simon, 2013), it would be likely that the response to the speech in noise would be delayed relative to the speech in quiet response. In terms of the scalp regions which most contribute to the first component, we found it to be dominated by central and temporal regions for speech in quiet, and for speech in noise there is a greater contribution from more parietal and occipital regions. This is likely due to increased contributions from the visual areas when the acoustic speech is noisy.

Limitations and future considerations. The use of CCA to study responses to natural speech has allowed us to answer questions that we could not previously answer. The use of natural and continuous stimuli is important in order to study the neural systems involved in processing audiovisual speech in the real-world (Hamilton & Huth, 2018). However, there are some drawbacks in the experiment design which could be improved upon in the future. The current paradigm is made to be somewhat unnatural by the presentation of auditory-only, visual-only and audiovisual speech, each in separate 1 minute trials. It is possible therefore, that in the visual-only condition, subjects find it very difficult to understand the speech and so may result in a decrease in attention to the speech material (or an increase for those subjects that are trying harder). If attention to the speech in visual-only trials differs in comparison with attention to the visual aspect of the audiovisual speech stimulus, then this could impact our visual-only EEG responses that are added to the auditory-only EEG responses to generate an A+V EEG response. This issue is common to almost all studies that examine multisensory integration effects of audiovisual speech (and indeed many multisensory experiments more generally).

One approach that has recently been suggested to overcome this is the use of audiovisual speech in every trial but varying the delay between the auditory and visual speech across trials such that the audiovisual speech is still intelligible. This variability in delay can theoretically allow one to characterize the unisensory responses using deconvolution (Metzger *et al.,* 2020). This approach was used for the presentation of single words but would be interesting to apply it in a paradigm using continuous speech.

Nevertheless, the effect of interest in this study is the relative change in the difference between the performance of an AV and A+V model applied to AV EEG responses across quiet and noisy speech conditions. We would not expect the difference in performance to change with different speech conditions if the improvement of the AV model was simply driven by a poor A+V model due to a poor contribution from the V-only response in the A+V EEG response.

4.5. Conclusion

This work has used a novel framework to study multisensory interactions at the acoustic and phonetic levels of speech processing. This has revealed that multisensory effects are present for both the spectrogram and phonetic feature representations when the speech is in quiet or when it is masked by noise. However, for speech in noise, multisensory interactions are significantly larger at the phonetic-feature level suggesting that in noisy conditions, the listener relies more on higher-level articulatory information from visual

speech.

Chapter 5.

The contribution of visual dynamic and articulatory cues to audiovisual speech perception in noise

In the previous chapter, we examined how visual speech impacts auditory speech at different stages of processing. We found that multisensory interactions were enhanced at the level of phonetic processing in degraded listening conditions. This effect may be driven by an increased reliance on visual articulatory information when the acoustics are noisy. In this chapter, we modulate the visibility of the mouth details (i.e., articulatory details) in order to probe how different aspects of visual speech information contribute to the multisensory interactions we have observed at different stages of auditory processing.

5.1. Introduction

As discussed in section 2.2, visual speech is thought to provide two forms of information that contribute to auditory speech processing, that of dynamic information which is correlated with the acoustic signal (Chandrasekaran *et al.*, 2009) and articulatory information which helps resolve phoneme identity (Summerfield, 1987; Campbell, 2008). Access to these cues results in improved speech intelligibility when the listener can both see and hear the talker in comparison with hearing alone (Sumby & Pollack, 1954; Reisberg *et al.*, 1987; Grant & Seitz, 2000; Ross *et al.*, 2007).

To better understand what aspects of the face provide these visual cues, studies have investigated the gaze behavior of people during a speech comprehension task. It was found that people spend the majority of time looking at either the mouth or the eyes of the speaker (Vatikiotis-Bateson *et al.*, 1998; Lansing & McConkie, 2003; Buchan *et al.*, 2007). This suggests that information related to the speech can be extracted from these regions. Efforts to isolate the information contained in different parts of the speaker's face have shown that the mouth, jaw, eyebrows and head all contain speech related information that can help comprehension (Summerfield, 1979; Jiang *et al.*, 2002; Yehia *et al.*, 2002; Munhall *et al.*, 2004; Thomas & Jordan, 2004; Davis & Kim, 2006). Head movements that occur as a person speaks have been shown to correlate with variation in speech fundamental frequency (F0) and amplitude (Munhall *et al.*, 2004), and between 80 and 90% of the variance observed in face motion can be accounted for by the speech acoustics (Yehia *et al.*, 2002). While much of this visual information may be redundant when one can hear the speech perfectly, when the listening conditions are challenging there is a

substantial improvement in comprehension obtained from visual speech (Ross *et al.*, 2007). In this case, people tend to increase the time spent looking at the mouth of the speaker (Vatikiotis-Bateson *et al.*, 1998; Buchan *et al.*, 2008) – and it was shown that the amount of mouth viewing time is positively correlated with speech comprehension (Bidelman *et al.*, 2020). This suggests that cues from the mouth provide information that directly affects speech comprehension.

Neuroimaging studies examining the role of visual cues in speech perception have also found the mouth area to be important - recent ECoG work showed that when watching silent visual speech the frontal cortex specifically enhances activity in the retinotopic area of visual cortex that encodes the mouth (Ozker *et al.*, 2018). In addition, pSTS has been shown to preferentially respond to mouth movements over other facial movements (Zhu & Beauchamp, 2017), as well as showing increased connectivity with visual cortex when the audio speech is noisy (Nath & Beauchamp, 2011). This finding is particularly interesting given that STG (which is directly adjacent to STS) has also been implicated in auditory phonetic processing (Mesgarani *et al.*, 2014) and is an important locus in the speech processing hierarchy (for review see: Yi *et al.*, 2019). Together, the evidence points to pSTS as a site where vocal sounds are integrated with information from the mouth - highlighting the importance of the mouth in impacting audiovisual speech perception.

It remains unknown however, what information from the mouth is used since the mouth contains dynamic as well as articulatory information. Specifically, it is unclear how the mouth dynamics and mouth details (e.g., teeth and tongue) differentially contribute to the perception of audiovisual speech.

One perspective of audiovisual speech integration posits that dynamic information from visual speech is projected directly from visual cortex to benefit audio speech processing at an early stage, i.e., in core sensory cortex. Then in the later stages of processing, the left TVSA region of visual cortex is thought to be involved in computing visual phonetic information which is then fed to the STS/G where phoneme identity is resolved (Peelle & Sommers, 2015). This framework attempts to account for how dynamic and articulatory information from the visual speech differently impact auditory speech processing. However, direct evidence to support this framework is lacking.

In the previous study, we showed that multisensory effects could be captured at both stages of processing (i.e., at the level of spectral and phoneme processing), however, in that study participants had access to both the dynamic visual information as well as

articulatory information. In this study we aim to disentangle these two components of visual speech by controlling access to the dynamic and articulatory information from the face of the speaker. The designed experiment blocks the articulatory information from the lips - while the mouth dynamics and the rest of the facial information remains relatively intact as participants listen to audiovisual (AV) speech in noise (-7 dB).

5.2. Material and Methods

The methodology used here is similar to Study 2 (Chapter 4.2).

Participants. Fifteen native English speakers (3 males; age range 18–40 years, mean age 24 years) took part in the experiment. All participants were free of neurological diseases, had self-reported normal hearing, and had normal or corrected-to-normal vision. Informed consent was obtained from all participants before the experiment, and they received monetary compensation for their time. The study was approved by the Research Subjects Review Board at the University of Rochester.

Stimuli. The speech stimuli were drawn from a collection of videos featuring a trained male speaker. The videos consisted of the speaker's head centered in the frame with a black background. Adobe After Effects software was used to perform detailed face tracking on each of the fifteen 1-minute videos. This provided tracking points on the circumference of the head, the eyes, nose and mouth. The tracking was manually checked to ensure accuracy. There were 4 tracking points on the mouth and these points tracked the top, bottom, left and right parts of the mouth. These points were used to generate dynamic masks for the blurred mouth conditions. The mouth tracking points were applied to a mask with a Gaussian blur (see fig. 5.1) which resulted in removing the visibility of any mouth details from the video (i.e., the teeth, lips and tongue) while retaining the dynamic mouth area information. This was done to generate the AV-blur and V-blur videos. The AV-clear and V-clear and A-only videos went through the same processing steps but no masks were applied.

Videos were rendered into 1920x1080 pixel movies with a frame rate of 30 fps in Adobe Media Encoder using the Xvid MPEG-4 video encoder. The soundtracks from each video were mixed with spectrally matched stationary noise generated using a 50th-order linear predictive model estimated from the original speech recording (Ding and Simon, 2013; Ding et al., 2013). Prediction order was calculated based on the sampling rate of the soundtracks (Parsons, 1987). This resulted in speech with SNR of -7 dB. The speech plus noise audio, with sampling rate of 48 kHz and 16-bit resolution, was then combined with each of the corresponding videos in VirtualDub (www.virtualdub.org). The stimuli used

in this study will be made available upon reasonable request to the author or advisor.

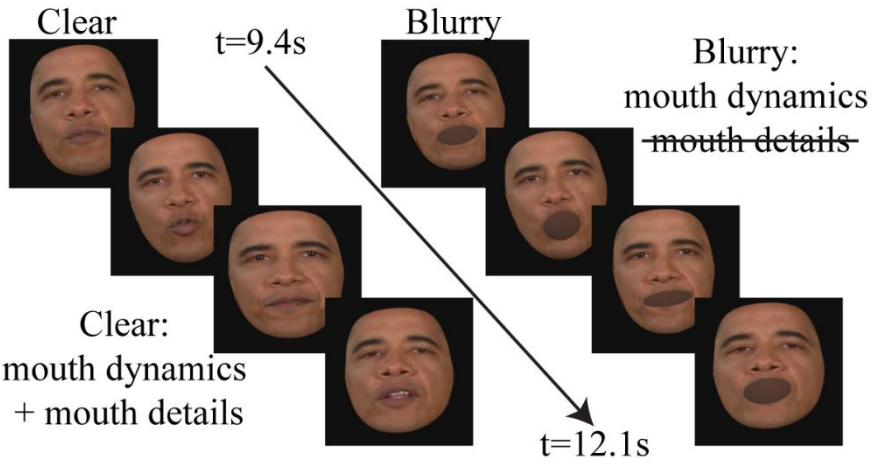


Figure 5.1. A 3-second extract of a video in the clear condition with the matching frames shown for blurred condition on the right. For the clear video, all visual speech features are available, whereas in the blurred condition the mouth details are removed and so subjects only have access to face dynamics, mouth dynamics and mouth area information.

Procedure. Stimulus presentation and data recording took place in a dark sound attenuated room with participants seated at a distance of 60 cm from the visual display. Visual stimuli were presented on a 26 inch LCD monitor operating at a refresh rate of 60 Hz. The face of the speaker subtended a visual angle of 15°. Audio stimuli were presented diotically through Sennheiser HD650 headphones at a comfortable level. Stimulus presentation was controlled using Presentation software (Neurobehavioral Systems). During the experiment, each of the 15 speech passages was presented five times, each time as part of a different experimental condition. There were 5 conditions in the experiment: (1) audiovisual speech with a clear face (AVclear), (2) audiovisual speech with a blurred mouth (AVblur), (3) auditory-only speech (A-only), (4) visual-only speech with a clear mouth (Vclear) and (5) visual-only speech with a blurred mouth (Vblur). These 5 conditions were used with the goal of comparing multisensory effects for the clear mouth condition (AVclear ~= (A+Vclear)) and multisensory effects for the blurred mouth condition (AVblur ~= (A+Vblur)). Presentation order was pseudo-randomized across conditions, ensuring that repeats were spaced at least 15 minutes apart for each participant. Participants were allowed free viewing of the stimuli, except for the audio only condition where they were instructed to fixate on the crosshair. They were also instructed to minimize eye blinking and all other motor activity during recording. Eye tracking was recorded for the duration of the experiment using the Tobii Pro X3-120 eye tracker with 120 Hz sampling rate. A chin rest was use to stabilize the participants head. Eye tracking data was recorded along with presentation event information using lab streaming layer (SCCN, San Diego) in order to synchronize the EEG with the eye data.

To encourage active engagement with the content of the speech, participants were required to respond to target words via button press. Before each trial, a target word was displayed on the monitor until the participant was ready to begin. A target word was deemed have been correctly detected if subjects responded by button press within 0–2 seconds after target word onset. In addition to detecting target words, participants were required to rate their intelligibility of the speech stimuli at the end of each 60-s trial. Intelligibility was rated as a percentage of the total words understood using a 10-point scale (0–10%, 10–20%, ... 90–100%).

EEG acquisition and preprocessing. The EEG data were recorded using an ActiveTwo system (BioSemi) from 128 scalp electrodes and two mastoid electrodes at a sampling rate of 512 Hz. Triggers indicating the start of each trial were recorded along with the EEG. Subsequent preprocessing was conducted off-line in MATLAB. The data were then bandpass filtered using second-order, zero phase-shift Butterworth filters between 0.3-30 Hz, downsampled to 64 Hz, and re-referenced to the average of the mastoid channels. Channels contaminated by noise were recalculated by spline-interpolating the surrounding clean channels in EEGLAB (Delorme & Makeig, 2004).

EEG analysis of multisensory speech processing. Our aim was to examine how the available visual information affects participants' understanding of the speech as well as how it effects our measures of the neural responses to the speech. In order to examine how the difference in visual information across conditions would impact the neural processing of auditory speech at different hierarchical levels, we used the same approach as in the previous chapter which involved calculating a spectrogram and phonetic feature representation of the speech signal and relating these representations to the EEG using CCA. See the subsections of the methods entitled '**Indexing neurophysiological speech processing at different hierarchical levels**', '**Canonical correlation analysis**' and '**Indexing multisensory integration using CCA**' for a detailed description of how CCA was performed with the spectrogram and phonetic features to index multisensory effects. Again, this approach was used here to examine the influence of the mouth blurring on tracking of the speech at different stages as well as to compare multisensory effects at these stages across the clear and blurred mouth conditions.

To relate our study to previous work on this topic we also looked at relating the EEG to the speech envelope. The envelope was calculated by first bandpass filtering the stimuli into 128 logarithmically-spaced frequency bands between 80 and 3000 Hz using a compressive gammachirp auditory filter bank that modeled the auditory periphery (Irino

& Patterson, 2006). The energy in each frequency band was calculated using a Hilbert transform and the broadband envelope was obtained by averaging across the frequency bands of the resulting spectrogram.

Multivariate regression is the analysis approach used here to relate the EEG to the envelope, which is the same as that used by (Crosse *et al.*, 2015a; Crosse *et al.*, 2016b). Using this approach we can index multisensory interactions based on the reconstruction of the acoustic envelope. To do this, we first measure the neural tracking of the speech signal in terms of how accurately the broadband speech envelope, *s*(*t*), could be reconstructed from the EEG data, *r*(*t*), using the following linear model:

$$\hat{s}(t) = \sum_{n=1}^{128} \sum_{\tau=0}^{500ms} r(t+\tau, n) g(\tau, n)$$

where $\hat{s}(t)$ is the estimated speech envelope, $r(t+\tau, n)$ is the EEG response at channel *n* and time lag τ, and $g(\tau, n)$ is the linear decoder for the corresponding channel and time lag. We used time lags of 0-500ms. The decoder $g(\tau, n)$ was optimized for each condition using ridge regression with leave-one-out cross-validation. The neural measure of multisensory integration was obtained by summing the unisensory A and V models and testing their ability to reconstruct an estimate of the speech envelope from the AV EEG data. The performance of the A+V model is then compared with the AV model, whereby any 'gains' achieved by the AV model are inferred to be caused by multisensory effects. The equation for measuring the multisensory integration effect is:

$$MSI_{Gain} = \frac{corr[\hat{s}_{AV}(t), s(t)] - corr[\hat{s}_{A+V}(t), s(t)]}{|corr[\hat{s}_{AV}(t), s(t)]| + |corr[\hat{s}_{A+V}(t), s(t)]|}$$

where $\hat{s}_{AV}(t)$ is the reconstructed envelope for the AV condition and $\hat{s}_{A+V}(t)$ is the estimated envelope for the additive unisensory model. The difference in performance between the models is normalized to ensure that the measure of multisensory integration is not be biased by different correlations across conditions as well as the variability in correlation values across subjects. Here, we had conditions of AV clear mouth and AV blurred mouth and generated the corresponding unisensory models for both (i.e., A+Vclear and A+Vblur). This allowed us to estimate multisensory effects for both the clear mouth condition and the blurred mouth conditions which we were interested in comparing with one another.

The linear decoder used for reconstructing the envelope, integrates information in the EEG data across time-lags of 0-500 ms. In order to examine the contribution of each time-

lag to the envelope reconstruction, we trained a decoder for each individual time-lag separately. This produces a reconstruction accuracy of the envelope at each time-delay. We can then use these reconstruction values per time-lag to calculate how multisensory effects varied as a function of time-lags by comparing the AV decoders' performance with the A+V decoders' performance.

Statistical analysis. All statistical comparisons were conducted using non-parametric permutation tests with 10,000 repetitions of randomly assigning the values from the two groups being compared and calculating the difference between the groups (Combrisson & Jerbi, 2015). Then the tail of this empirical distribution is used to calculate the p-value for the actual data, and two-tailed tests are used throughout. Where multiple comparisons were carried out p-values were corrected using the False Discovery Rate (FDR) method (Benjamini & Hochberg, 1995). All numerical values are reported as mean ± SD.

5.3. Results

The data shown here are for 14 out of the 15 subjects recorded due to 1 subject detecting no targets across all conditions and so for all of the analyses we used the remaining 14 subjects. Removing the mouth detail from the videos resulted in a decrease in understanding of the speech material reflected in both target word detection performance and intelligibility rating (Fig. 5.2A, B; $p<0.0001$ for both). We also found significant positive correlations between target word detection and intelligibility rating for each condition containing sound (AVclear, $r = 0.56$, $p=0.03$, AVblur, $r=0.82$, $p=2.9x10^{-4}$, A-only, $r=0.66$, $p=0.009$, Fig. 5.2C). For the video only conditions, there was no correlation between the two behavioural measures, likely due to the fact that the intelligibility rating was close to floor level (Vclear, $r=0.23$, $p=0.42$, Vblur, $r=-0.22$, $p=0.44$, Fig. 5.2C).

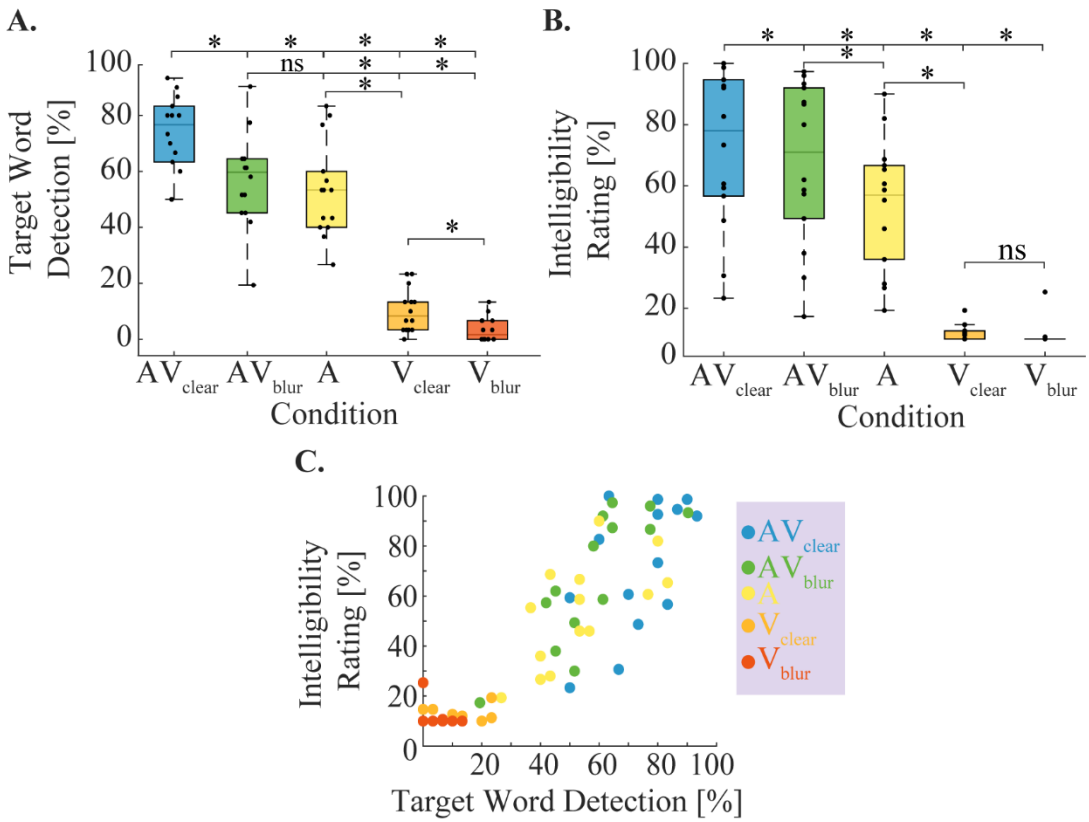


*Figure 5.2. Behavioural performance for each condition. A. Target word detection performance. B. Self-rated intelligibility scores of the speech. * indicate p<0.05. C. Correlation between the two behavioural measures is strong and positive (p<0.05) for all conditions except the visual-only conditions (p>0.05) due to performance being close to floor level.*

We then examined the eye-tracking data across the different conditions in order to assess if the reduction in behavioural performance may have been driven by different viewing behaviours across conditions. There was one subject for which data was not obtained due to technical issues, and so the eye tracking data is for 13 out of the 14 subjects. To compare viewing behaviours across conditions, an area around the mouth was defined as the mouth ROI (see Fig. 5.3A, top left), and the percentage of time each subject was recorded looking at that area was termed the percentage of mouth viewing time. This showed that there was a significant reduction in the amount of time participants spent looking at the mouth in the conditions where the mouth was blurred compared with the conditions where the mouth was clear (AVclear > AVblur, p<0.0001, Vclear > Vblur, , p<0.0001). The result for the A-only condition is plotted here also for completeness, although there was no visible face in the A-only condition and instead a crosshair was present at the centre of the screen, at which subjects were asked to fixate on.

Heatmaps of subjects' gaze for the clear and blurry conditions show somewhat similar viewing behaviour which was predominantly focused on the mouth in both conditions, however in the blurry condition there is a decrease in time spent looking at mouth (Fig.

5.3B, p<0.0001). This behaviour is likely due to a reduction in information available from the mouth resulting in a need to extract additional information from the face.

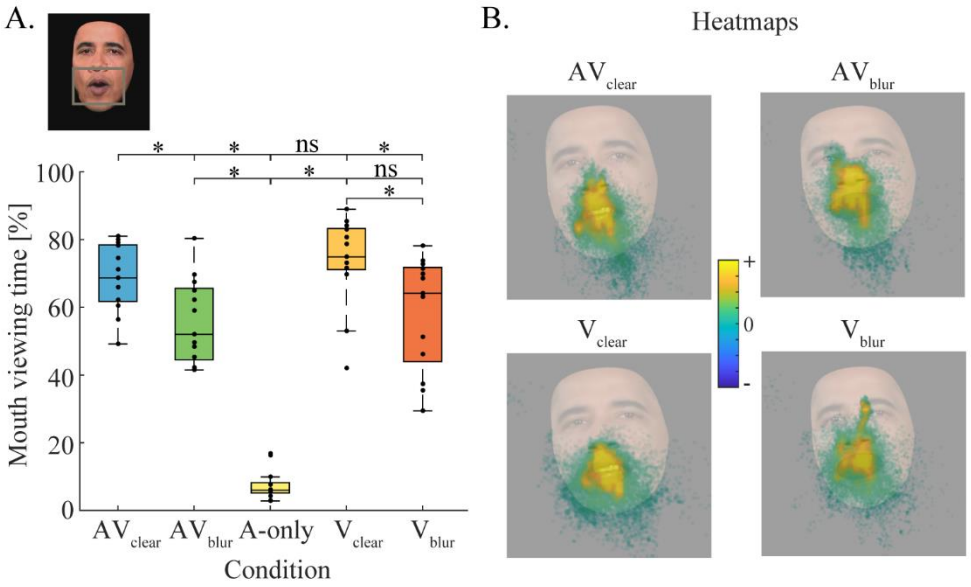


Figure 5.3. Gaze behaviour of subjects for each condition. A. There is a significant reduction in mouth viewing time for the blurred condition compared with the clear mouth condition. B. Heatmaps show somewhat similar viewing behaviors across conditions with a reduction in mouth viewing time for the blurred mouth conditions.

We then examined if there was a relationship between how much time subjects spent looking at the mouth and their understanding of the speech, as has been reported previously (Bidelman *et al.*, 2020). While we found a positive trend between the amount of mouth viewing time and the behavioural performance, it was only significant for the self-rated intelligibility of the speech in the blurry condition.

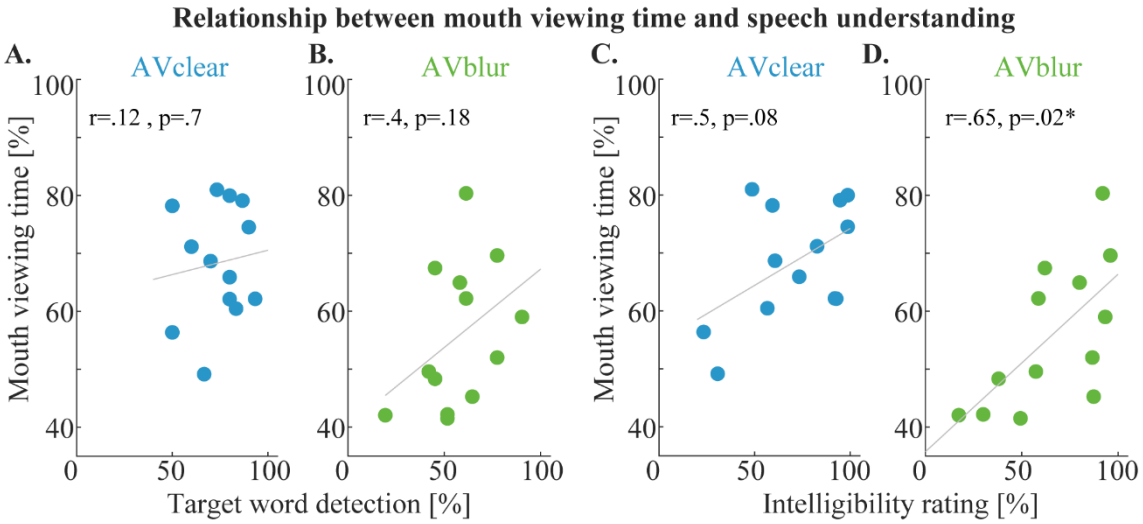


Figure 5.4. Relationship between speech understanding and amount of time spent looking at the mouth. A. Correlation between target word detection and the mouth viewing time in the AVclear condition. B. Correlation between target word detection and the mouth viewing time in the AVblur condition. C. Correlation between self-rated speech intelligibility and the mouth viewing time in the AVclear condition. D. Correlation

between self-rated speech intelligibility and the mouth viewing time in the AVblur condition. This is the only condition for which there is a significant positive correlation between mouth viewing time and behaviour.

We then wanted to investigate if there are underlying differences in neural responses that give rise to this reduction in speech understanding when the mouth details are blurred. Firstly, we performed envelope reconstruction to test how the speech envelope encoding was affected by blurring the mouth. The goal of this analysis was to assess whether blurring the mouth details would impact envelope tracking and what if any would be the effect on multisensory interactions. It was not clear whether blurring the mouth would impact envelope tracking given that it is possible that mouth dynamics may be all that is reflected in the envelope tracking measure. However, the speech envelope is correlated with both acoustic and linguistic representations of the speech signal making it more difficult to make interpretations from results based on envelope tracking alone. Nonetheless, it has been the most prominent representation used in studies of speech processing (Ding & Simon, 2014; Obleser & Kayser, 2019) and more recently has been used in studies of audiovisual speech processing (Golumbic *et al.*, 2013b; Crosse *et al.*, 2015a; Crosse *et al.*, 2016b) and so it allows us to more directly compare our results with earlier work on this topic.

Our results showed a small difference in reconstruction accuracy between the clear and blurry AV conditions, resulting in a weak effect (Fig. 5.5A, $p=0.08$). Comparing the V-only conditions, we found significantly greater reconstruction accuracy for the Vclear condition compared with the Vblur condition (Fig. 5.5A, $p<0.0001$). We then used our framework for identifying multisensory integration effects in EEG in order to assess differences in multisensory integration across the 2 conditions. We found significant multisensory effects for both clear and blurry speech (AV>(A+V), $p<0.0001$ and $p=0.002$ respectively), however, there was no difference in the multisensory gain across the 2 conditions (Fig. 5.5B, $p=0.3$).

In order to examine the cortical regions that contributed to each of our decoders we plotted the forward transformations of the decoder weights, which are more interpretable in terms of the underlying physiology (Haufe *et al.*, 2014). As shown in Fig. 5.5C, the AVclear and AVblur decoders show strong activations over temporal scalp bilaterally, as well as activations over occipital scalp, suggesting the involvement of auditory and visual cortices. The A-only decoder shows bilateral temporal activations while the Vclear and Vblur decoders have contributions from occipital regions (Fig. 5.5C).

We then examined the reconstruction accuracy at the level of single lags in order to

examine the timecourse of the model performances. This revealed significant multisensory effects for the clear mouth condition between -50-200ms (Fig. 5.5D). However, for the blurry mouth condition there was no timelag at which there was a significant difference between AV and A+V decoders (Fig. 5.5E). This result is in contrast with the positive multisensory gain found for 11/14 subjects in the AVblur condition using the multi-lag reconstruction model. This may be due to the fact that the single lag model is not as sensitive as the multi-lag model, due to the ability of the multi-lag model to integrate across time-lags. Nevertheless, the single-lag reconstruction approach shows robust multisensory effects for the clear mouth condition while there is no apparent multisensory effects for the blurry mouth condition. This result is in line with our hypothesis that multisensory integration is enhanced when mouth details are visible.

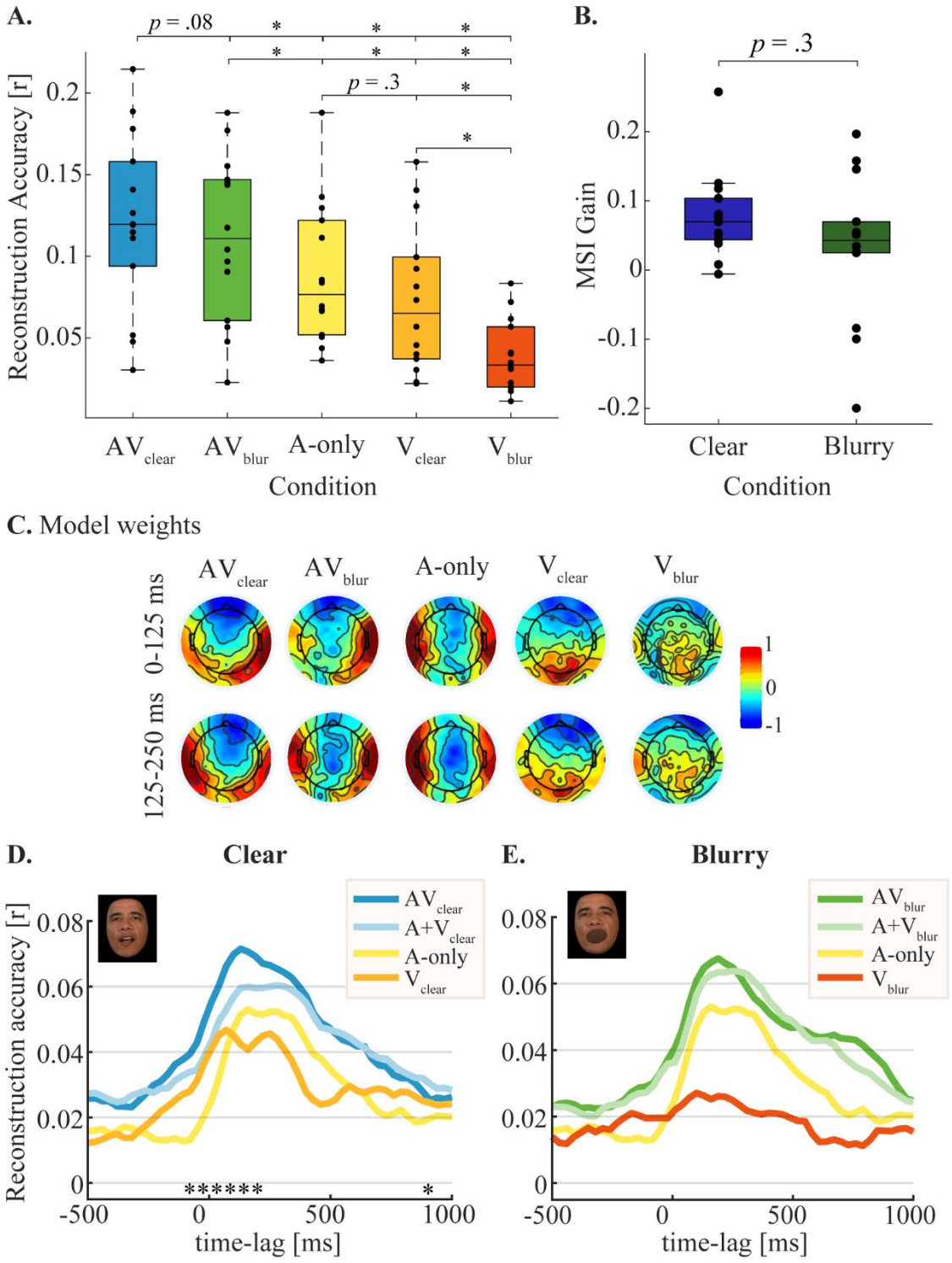


Figure 5.5. Envelope reconstruction results. A. Multi-lag envelope reconstruction results for all conditions. The AVclear condition has the highest reconstruction accuracy although it is similar to the AVblur condition (p=0.08). B. Multisensory effects for clear and blurry conditions show that there is no difference in multisensory integration effects across the 2 conditions. C. The model weights are transformed into forward weights for each condition. D. Single-lag reconstruction accuracy for the clear mouth conditions, this reveals multisensory effects present at -50-200 ms lags. E. The same plot for the blurry mouth condition, in this case there is no time-lag at which there is a significant difference between AV and A+V.

While the envelope is a useful measure for assessing the cortical tracking of speech, it

necessarily conflates multiple levels of speech processing, making it difficult to interpret results from this analysis. Since we are interested in assessing the impact of the mouth blurring on different stages of speech processing, we wanted to examine the cortical tracking of the spectrogram and phonetic feature representations of the speech signal using CCA, as we did in the previous chapter. Again, we partialled out the phonetic features from the EEG using the TRF method and used the residual EEG to relate it to the spectrogram representation using CCA. Similarly, for examining the relationship between the EEG and the phonetic features, we used EEG which already had the spectrogram regressed out. This allowed us to examine the unique contribution of each of these representations. This approach was previously described in more detail in chapter 4 (see section 4.3, subsection entitled '**Isolating multisensory effects at the spectrotemporal and phonetic levels**').

Relating the unique spectrogram representation to the EEG (phonetic features partialled out) we found multisensory effects for both clear ($p=0.003$) and blurry mouth conditions ($p=0.003$). However, there was no difference in performance between the AV clear and AV blur models (Fig. 5.6A; $p=0.7$), suggesting that spectrotemporal information was tracked to a reasonably similar extent in both conditions. Comparing the multisensory gain across the 2 conditions we also found no difference in gain (Fig. 5.6B; $p=0.9$).

Using the phonetic feature representation (spectrogram partialled out), we found multisensory effects (AV>A+V) for both the clear mouth ($p=0.01$) and the blurry mouth conditions ($p=0.008$). We also found a significant difference between the AV clear versus AV blurry correlations (Fig 5.6C; $p<0.0001$). This suggests that there is some difference in phonetic tracking across the clear and blurred mouth conditions, which we did not see in the case of the spectrogram representation. When we calculated the multisensory gain however, we found no difference in gain across conditions (Fig. 5.6D; $p=0.9$). The lack of a difference in multisensory gain here is surprising as our original hypothesis was that multisensory effects would be reduced at the level of phonetic processing when the mouth details are blurred.

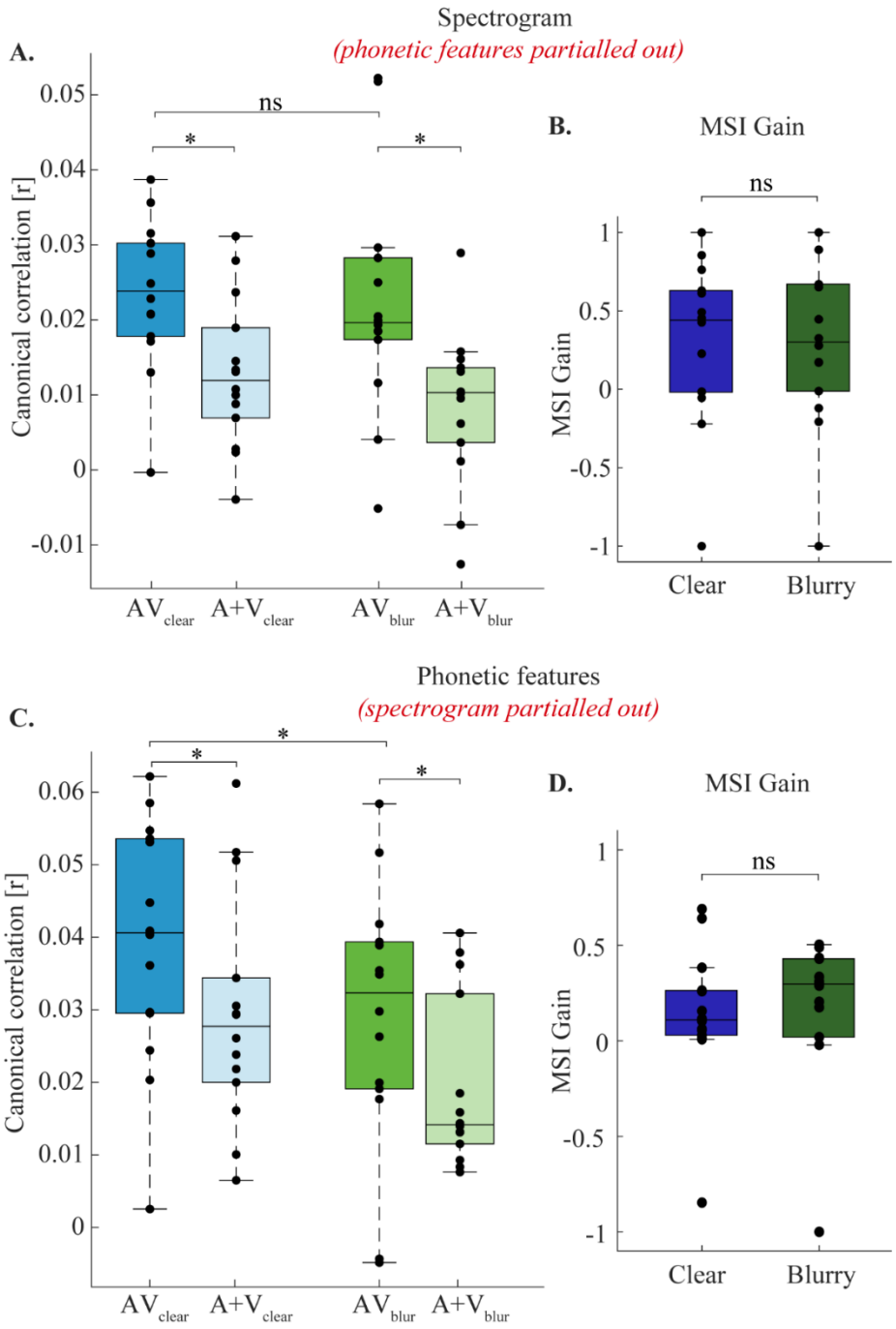


Figure. 5.7. Correlations for the first canonical component for spectrogram and phonetic feature representations of the speech signal in clear and blurred mouth conditions. A. The correlation of the first canonical component using the spectrogram representation of the speech signal and the EEG that has phonetic features regressed out. Multisensory effects were found for both clear and blurred mouth conditions (p=0.003 for both). B. The multisensory gain for clear and blurry conditions showed a majority of positive values of gain for both conditions but there was no difference between the conditions. C. The correlation of the first canonical component using the phonetic representation of the speech signal and the EEG that has the spectrogram regressed out. Multisensory effects were found for both clear and blurred mouth conditions (p=0.01 and p=0.008). D. Multisensory gains for both conditions using calculated from the phonetic feature correlations also showed no difference in gains.

5.4. Discussion

In this study, we have investigated how mouth dynamics and mouth details differentially contribute to the understanding and the neural representation of audiovisual speech in noise. Our results show that both clear and blurred mouth conditions have similar multisensory gain values when using the multi-lag envelope reconstruction model. However, there were significant multisensory integration effects (AV>A+V) at the level of single time-lags for the clear mouth condition, but no integration effects for the blurred mouth condition. Using the spectrogram and phonetic representations of the speech, we find an enhanced representation of phonetic features for the AVclear mouth compared with the AVblur condition (Fig. 5.6C), whereas for the spectrogram representation there is no difference in performance across conditions (Fig. 5.6A). These results provide tentative evidence that multisensory integration of AV speech is enhanced by the visibility of mouth details, and that this is driven by improved encoding of phonetic features.

Blurring the mouth details impairs speech understanding and reduces the time spent looking at the mouth. It is well known that the mouth provides a substantial amount of information related to the acoustic speech signal and can help disambiguate acoustically similar phonemes (Summerfield, 1979; Reisberg *et al.*, 1987), since any particular mouth movement is associated with a small number of auditory phonemes (Neti *et al.*, 2000; Cappelletta & Harte, 2012). However, the relative contribution of mouth details and mouth dynamics to the improvement in understanding of AV speech in noise remains unclear. We found that removal of the mouth details resulted in a decrease in speech understanding. This was true for both the target word detection and subjectively rated intelligibility measures. This supports the idea that mouth details provide additional information about the speech content above and beyond what is understood from using dynamics alone.

Our finding of a reduction in time spent looking at the mouth when the mouth details were blurred is in line with previous work that found a reduction in the amount of mouth fixations when the visual resolution was decreased (Wilson *et al.*, 2016). The reduction in mouth viewing time observed here, may partly explain the decrease in speech comprehension, since we found mouth viewing time to be positively correlated with intelligibility rating of the speech for the blurred mouth condition. One possible explanation for the reduction in mouth viewing time in the blurred mouth condition is that there was less information available from the mouth in comparison with the clear

condition. Viewers may have sought to compensate for this reduction in information by looking elsewhere on the face to extract other speech related cues. It is not clear however, to what extent the visual information provided by the rest of the face (outside of the mouth region) contributed to the understanding of the speech.

Blurring the mouth details results in reduced multisensory effects for AV speech in noise at the level of single-lags. The second stage of the multistage integration framework refers to the constraining of lexical competition that is contributed to by phonetic information (Peelle & Sommers, 2015). Identification of the lexical item is dependent on the phonemes or phonetic features which have been detected in the audiovisual speech. The mouth details are thought to contribute to this stage by providing information on place and manner of articulation which are critical for identifying consonants and provides information that is complementary to the acoustics (Walden *et al.*, 1977). Therefore, it was expected that removal of the mouth details would impact the second stage of integration. This was hypothesised to lead to an overall reduction in the measured multisensory integration effects, driven by a reduction in this second stage of integration. Performing envelope reconstruction at individual lags, we found significant multisensory interactions (AV>A+V) for the clear mouth condition from -50-200 ms, this is similar to what has been shown in previous work examining multisensory effects of speech in noise (Crosse *et al.*, 2016b). This result is in contrast with the blurred mouth condition, where at the level of single lags there is no difference between the AV and A+V decoders. This result suggests a reduction in multisensory effects for the blurred mouth condition. However, when we examine the multisensory gain for the multi-lag envelope reconstruction model, we do not find any difference in gain between the clear and blurred mouth conditions. Together, this suggests that multisensory effects are present in the case of the blurred mouth condition, but appear reduced such that they are not detectable at the level of single lags. The envelope representation of the speech is not readily interpretable in the context of speech processing since it is a simple representation of the amplitude of the sound waveform. In order to investigate multisensory effects at early and late stages of speech processing we examined the encoding of the spectrogram (acoustic) and phonetic features (speech specific) representation of the speech.

Blurring the mouth details negatively impacts phonetic encoding of the speech. It has been shown for natural speech that changes in mouth area are correlated with the modulation of the speech sound as well as with the formant frequencies (Stevens & House, 1955; Summerfield, 1992). Also, speech-based shape deformations are tightly

linked with the frequency content of the speech and have been shown to specifically enhance understanding of spectrally degraded speech (Plass *et al.*, 2019). Thus, much of the visual information that relates to the spectrotemporal content of the speech signal was expected to be available in the blurred mouth condition. In line with this, we found that there was no effect of blurring the mouth on the encoding of the spectrogram representation of the speech. There was also no difference in multisensory effects for the spectrogram in the clear and the blurred mouth conditions. This suggests that mouth details do not contribute unique information to the spectral processing of the speech.

In contrast with the results using the spectrogram representation, we expected that blurring of the mouth details would have significant impact on phonetic encoding of the speech due to the lack of availability of visual phonetic information. In line with this, we found there was a significant decrease in performance of the phonetic model when the mouth details were blurred. This suggests that at the level of phonetic feature processing, there is a significant reduction in encoding when the mouth details are removed. This is in line with our hypothesis that blurring the mouth details would result in removal of visual articulatory information which would have a more detrimental effect on phonetic processing than spectral processing.

We also hypothesised that multisensory integration effects at the stage of phonetic processing would be markedly reduced. However we found no difference in multisensory gain for phonetic features across the 2 conditions. Previous work which involved blurring the speakers face using spatial quantisation (i.e., local averaging of pixels into larger blocks) has suggested that visibility of the mouth details is not required to provide visual phonetic details (MacDonald *et al.*, 2000). This was based on findings that when the level of spatial quantisation was high (11.2 pixels/face) such that mouth details were no longer clear, visual phonetic information (e.g., syllable ‘ga’) “fused” with auditory phonetic information (e.g., syllable ‘ba’) such that subjects reported perceiving an illusory syllable (e.g., ‘da’). However, it is unknown how this finding would translate to a paradigm using natural speech, and in particular to what extent visual phonetic information is retrievable from areas of the face outside of the mouth (Vatikiotis-Bateson & Munhall, 2015).

In future work it would be of interest to tease apart the specific contributions of the face details and head dynamics from the mouth information, in order to better understand the visual speech cues provided by the rest of the face. This could be done by blurring the face such that details of movements of the jaw and cheeks and eyes would be removed while preserving the overall shape and dynamics of the face. The current study also

suffers from a relatively small sample size (n=14), and it may be necessary to collect data from over 20 participants in order to have enough statistical power to detect differences in multisensory integration effects.

5.5. Conclusion

In summary, we have examined the effect of mouth details on speech understanding, viewing behaviour and the underlying electrophysiological responses. We found that removing the mouth details results in a decrease in speech understanding and also a reduction in time spent viewing the mouth. This is likely driven by a reduction in information available from the mouth. The EEG analysis using envelope reconstruction revealed that at the level of single time-lags, there are significant multisensory integration effects for the clear mouth condition, whereas there are no integration effects for the blurred condition. This suggests a reduction in multisensory integration effects when the mouth details are blurred. In an attempt to isolate at which stage of speech processing the removal of mouth details most affects, we used CCA to examine the performance of a spectrogram and phonetic feature representation of the speech. This showed that there was a reduction in phonetic feature processing for the blurred mouth condition compared with the clear condition. There was no difference for the case of the spectrogram representation. This suggests that blurring the mouth details specifically influenced phonetic level processing in line with the idea that mouth articulatory details contribute to phonetic level processing of the speech. However, it is not clear from this analysis whether this reduction in performance is due to a lack of a visual phonetic representation in visual cortex or whether there is a reduction in auditory phonetic processing due to a decrease in speech understanding. As discussed in section 2.2.4, it is debated whether visual cortex processes visual speech as speech or whether it is auditory cortex and association cortex which carries out these computations on the visual speech.

The finding of higher reconstruction accuracy for the Vclear condition compared with the Vblur condition is also quite interesting and is worthy of further exploration. This points to an enhancement in visual speech processing in visual regions (evident from the activation patterns in Fig. 5.5C) when the mouth details are visible. In this study, we were particularly interested in how this increase in visual information impacts auditory processing, however, based on our results we cannot make any conclusive statements on this.

Chapter 6.

General Discussion

Speech in its natural form is multisensory, yet the role that visual speech plays in our perception of speech is not well understood. In this thesis, we used EEG and multivariate analysis techniques to examine this question in the context of natural audiovisual speech.

Using multivariate regression in chapter 3, we found that when subjects pay attention to auditory speech with matching visual speech, the neural activation patterns are different from the activations when subjects attend to auditory speech in the presence of irrelevant visual speech. The differences in activations are likely driven by the presence of multisensory interactions when the auditory and visual speech are matching, whereas these interactions are not present when the visual speech is not matching. The finding of a substantial reduction in parieto-occipital alpha power when the visual speech is relevant compared with when the visual speech is irrelevant, also suggests an increased contribution from visual cortical regions when processing the relevant visual speech. This paradigm however, did not allow us to separate contributions from visual cortical regions and multisensory interactions. In the subsequent studies we remedied this by using paradigms which presented auditory-only, visual-only and audiovisual speech in separate conditions. This enables one to account for the expected unisensory processing using the combination of auditory-only and visual-only responses and compare this with the response to the multisensory stimulus (i.e., audiovisual).

6.1. Insights on current models of audiovisual speech perception and integration

Multisensory integration in the more general sense, i.e., not just related to speech processing, has previously been described as following the principle of inverse effectiveness. This refers to the idea that the largest multisensory enhancement is expected when a unisensory stimulus is weakest (Meredith & Stein, 1986). In other words, when one modality of the stimulus is degraded, the percept is enhanced by the presence of a second modality, or if both modalities are degraded then the enhancement is even greater, and so they are inversely effective. However, in the case of speech processing it was found that this principle does not explain behaviour (Ross *et al.*, 2007).

The finding that audiovisual speech integration does not follow the principle of inverse effectiveness motivated alternative descriptors. Bayesian cue integration is one

alternative that appears to describe audiovisual speech perception well (Ma *et al.*, 2009). This model takes into account the relative reliability of the auditory and visual inputs in order to determine the speech sound or word. This model also generalises well to other forms of cue integration and so is not thought to be unique to audiovisual speech integration (Ernst & Banks, 2002; Körding & Wolpert, 2004). While this model describes perception well, it does not tell us about the underlying neural responses to the speech and where they are integrated. Current models on neural mechanisms underlying the integration of auditory and visual speech suggests a multistage process as described in section 2.3.2. While this model presents a well-defined framework there is little evidence to directly support such a framework, particularly in the context of natural speech.

In chapter 4, we implemented a paradigm that allowed us to investigate and quantify multisensory integration effects at 2 different stages of speech processing. We found multisensory integration effects for both the spectral and phonetic stages when speech was presented without any background noise. The finding of integration effects at 2 stages of speech processing supports the idea of a multistage integration framework. We then showed that when background noise is added to the speech, there are also significant multisensory effects. In this case however, the integration effects at the level of phonetic processing are greater than for speech in quiet. This result points to a greater reliance on articulatory information when the acoustics are buried in noise and it is thought that visual articulatory details contributes most to this stage of processing. Although the spatial resolution of EEG does not allow us to state the sources of these effects, we would propose that, in line with the multistage model, the integration effects at the level of spectral processing occur in core regions of auditory cortex (Möttönen *et al.*, 2002; Okada *et al.*, 2013) and that the integration effects at the phonetic level occur in superior temporal cortex (Arnal *et al.*, 2009; Zhu & Beauchamp, 2017).

One property of audiovisual speech which we haven't examined here, and which has been shown to be a main driver of multisensory effects in pSTG, is the natural variation in the temporal relationship between the auditory and visual speech. In particular, a recent study using ECoG found that the neural response in pSTG was suppressed for audiovisual words where the visual speech was leading the auditory speech, relative to the response to auditory-only words (Karas *et al.*, 2019). There was no difference in the response however, when the visual speech was not leading the auditory speech. While this finding highlights that the natural temporal relationship between visual and auditory speech is an important contributor to multisensory effects, it also raises the question of what underlies

the perceptual benefit shown for audiovisual speech when visual speech is not leading auditory speech. It may be possible that the same mechanism explains the perceptual benefit of audiovisual speech, when the visual speech is not leading, but the analysis approach in this case might not have been sensitive enough to detect it. If noise was added to the auditory speech, then the perceptual benefit of audiovisual speech when the visual speech is not leading would be larger (compared with speech in quiet), which may provide greater power in the analysis to detect changes in the neural response for non-visual leading audiovisual speech.

Nevertheless, it would be of interest to characterise the natural delays between the auditory and visual speech for a continuous stream of speech (the abovementioned study used just 4 isolated words) and apply the framework used here to assess the impact of visual leading words versus other words on our measure of multisensory interactions. This approach could be insightful given the ability to use a much richer set of stimuli and also the ability to combine responses from the whole brain – the ECoG sensors usually only cover a small area of the cortical surface. If this result is reproduced under more naturalistic conditions, then it could have important implications for current models of multisensory integration and in particular how they should be modified to account for differing effects depending on the auditory-visual timing relationship.

6.2. The impact of visual articulatory detail on audiovisual speech processing and where this information is computed

The articulatory details of the mouth (i.e., the position and visibility of the lips, teeth and tongue) provide complementary information about the acoustic signal which is particularly important for understanding speech when listening conditions are degraded or when hearing is impaired. In chapter 5, we removed this information by blurring the mouth. This was done with the goal of impoverishing the visual speech of phonetic cues. It has previously been shown that the most salient visual phonetic cues are place and manner of articulation (Walden *et al.*, 1977) – features that are mostly characterised based on lips, teeth and tongue configurations.

We found that phonetic feature encoding was negatively impacted by blurring of the mouth, whereas there was no effect on encoding of spectral information. While this result was in line with our hypothesis, we also expected to see reduced multisensory effects at the level of phonetic processing which was not found. Similarly, using the envelope reconstruction approach, we did not find a difference in multisensory integration effects, except at the level of single time-lags. This analysis revealed multisensory effects for the

clear mouth condition across a range of lags (-50-200 ms) whereas there was no multisensory effects found for the blurred mouth condition. This result fits our hypothesis of a reduction in multisensory integration for the blurred mouth condition, but this result is weakened by the fact that we don't see the effect for the multi-lag model or for the phonetic feature model. This may be due to the fact that the expected effect size is small given that we expect most of the AV neural activity to be explained by the A+V model and only a small difference is expected to be related to multisensory effects. The sample size used in this study is also quite small (n=14), due to data acquisition stopping early because of coronavirus. It is likely that a sample size of 20+ people would be more appropriate and would allow for making stronger conclusions based on the results.

The SNR chosen for this experiment was based off a combination of pilot testing and previous work showing a multisensory 'sweet-spot' around this SNR for natural speech (Crosse *et al.*, 2016b). It is likely however, that the sweet-spot for individual subjects may vary and so a more formal approach to defining the optimal SNR for these experiments (possibly within subjects) would provide greater sensitivity for detecting multisensory effects, and for observing changes in multisensory integration based on the availability of the mouth details.

Although the objective of the chapter 5 study was to separate the contribution of dynamic and articulatory information, this is not a straightforward task since the mouth dynamics provides mouth area information which can help to disambiguate between high and low vowels (Plass *et al.*, 2019). It has also been shown that cheek and jaw movements are correlated with tongue movements (Yehia *et al.*, 1998; Jiang *et al.*, 2002). Therefore, it is possible that without seeing the mouth details, viewers can infer some of the information contained in the mouth details based on viewing other parts of the face. However, very little is known about the contribution of each part of the face to auditory speech processing. While it has been shown that people shift their gaze to the mouth more as the speech becomes noisier (Buchan *et al.*, 2008; Bidelman *et al.*, 2020), there is still a significant amount of time (about half) spent looking at the other parts of the face (Vatikiotis-Bateson *et al.*, 1998; Lansing & McConkie, 2003). Questions remain as to how the information extracted from these regions outside of the mouth are used – i.e., do they contribute to spectro-temporal or phonetic speech processing or both, and what role do they play in emotional and semantic processing of the speech.

The finding of a reduction in phonetic encoding for the blurred mouth condition in chapter 5 may be driven by a reduction in visual encoding of phonetic features or may stem from

a reduction in auditory encoding of phonetic features due to the impoverished visual information. This opens another area of research which was discussed in section 2.2.4, regarding the capability of visual cortex to carry out speech specific (phonetic) processing that it can then transmit to auditory (or heteromodal) regions. While there is a growing body of evidence for speech specific processing taking place in visual cortex (Bernstein *et al.*, 2011; Bernstein & Liebenthal, 2014; O'Sullivan *et al.*, 2017; Hauswald *et al.*, 2018), there is also recent evidence of dynamic tracking of silent speech in auditory regions. However, the auditory found tracking at a rate of 0.5 Hz (Bourguignon *et al.*, 2020). This is approximately the rate of sentences, and so does not explain where phonetic and lexical information is processed.

6.3. Future work

The main finding of chapter 4 of this thesis is that multisensory effects at the level of phonetic processing are enhanced when the acoustics are noisy. It is important to note however, the small effect size for this result. This weak effect is likely due to the fact that the conditions of quiet and noisy speech were carried out in 2 different groups of people, preventing the use of within-subject comparisons, and so making it more difficult to detect the expectedly small effects. Future work could design a within-subject experiment with eye-tracking that examines the effects of noise at each stage of speech processing. In this case, the effect size would be expected to be larger and furthermore, this design would enable the examination of changes in viewing behaviour across quiet and noisy speech conditions and how this impacts understanding as well as the underlying neural responses. This has the potential to provide a more well-informed understanding of multisensory integration in different environments, individual differences in viewing strategies and the flexibility of integration at different stages of speech processing.

In chapter 5, we examined the effect of blurring the mouth details on speech tracking and multisensory integration. However, the interpretation of the results from this study would benefit from a better understanding of the contribution of the rest of the face details (outside of the mouth area) to audiovisual speech perception. It would be particularly valuable to study how each aspect of the face contributes to speech processing across a range of different SNRs. The SNR at which each feature of the face contributes most to AV speech perception may differ, for e.g., one might expect that head movements correlated with the speech waveform would be beneficial at low SNRs whereas at higher SNRs the head dynamics may not provide much benefit to AV speech perception. On the other hand, mouth details that contain articulatory information would likely be helpful at

all SNRs but maximally beneficial at intermediate SNRs. This would provide a more comprehensive understanding of the contribution of facial features to auditory speech processing.

In this thesis, most of our results are presented in the context of the influence of visual speech on auditory speech tracking. However, it is difficult to capture 'purely' auditory speech features or similarly difficult to capture 'purely' visual speech features. This is due to the fact that these features are all necessarily correlated with each other given that the physical deformation of the face and vocal apparatus produce the acoustic speech waveform. While we have used partial regression to help account for correlations between features, an alternative approach which has recently emerged, known as back-to-back ridge regression, allows determination of the unique contribution of a whole array of auditory and visual speech features (King *et al.*, 2020). This could be useful for figuring out what features of the auditory and visual speech signals the brain encodes.

With an increase in features to be fit comes a necessary increase in data acquisition. This can be challenging for these types of multisensory experiments because they have 3 conditions (audio-only, visual-only and audiovisual). One novel approach which may remedy the burden of large amounts of data collection is deconvolution. This method was applied to data from iEEG from subjects presented with audiovisual speech where there were a range of time-delays between the auditory and visual speech onset times in order to assess the unisensory responses (Metzger *et al.*, 2020). The variable asynchrony between the auditory and visual speech onset times prevents collinearity of the auditory and visual responses. Therefore this approach allows the characterisation of unisensory responses through the presentation of audiovisual speech. An added benefit of this approach is that the variable delay is kept within a range of values such that the speech is still perceived as intelligible. This overcomes the issue of presenting visual-only speech where subjects often find it unintelligible, and so may be less likely to pay attention to the speech in this condition which could in turn affect the neural response.

Lastly, the term 'natural speech' has been referred to throughout this thesis, it is worth noting however, that the speech used in the experiments reported in this thesis are 'natural' in so far as it continuous and involves narratives that carry over from sentence to sentence without the need for repeated trials. However, to achieve truly natural speech, it would be necessary to design experiments whereby participants engage in face-to-face conversation while their EEG is recorded. This would be a step closer to true naturalness and may give a more realistic view on how people gaze at the face of the person they are

talking to in the real-world, rather than when the person is appearing on a pre-recorded video. It may also provide new insights on the neural tracking of speech in these circumstances.

6.4. Summary and Conclusions

This thesis addresses questions on how visual speech influences processing of natural, continuous auditory speech. In the first study, we used a regression-based framework to relate the EEG to the continuously changing speech envelope, when the visual speech was either matching the attended auditory speech or matching the ignored auditory speech. This revealed differences in neural activation patterns across the two conditions, and a suppression of visual processing when the visual speech did not match the attended auditory speech.

In the second study, we designed an experiment to specifically isolate multisensory interactions using a novel analysis technique. Using canonical correlation analysis we related multivariate representations of the speech to the multivariate EEG responses. This showed multisensory integration effects at the level of spectral and phonetic processing and showed that when the acoustics are noisy, multisensory effects are enhanced at the phonetic level of processing. In the third and final study, we removed mouth details from the visual speech using blurring while retaining mouth dynamics and all other facial features intact. We found that phonetic encoding was reduced when the mouth was blurred compared with when it was clear, while there was no effect on spectrogram encoding. This suggests that the mouth details contribute to phonetic encoding of the speech.

These studies highlight the substantial influence of visual speech on auditory speech processing in the brain. Beyond the well-known perceptual effects of visual speech, we show how visual speech impacts different stages of the auditory processing hierarchy. We also show that multisensory integration adapts to changes in the environment by enhancing integration at the phonetic level when the acoustics are degraded. Lastly, we show that the speech processing hierarchy is influenced by the available visual information from the face with a specific reduction in phonetic encoding when the mouth details are blurred.

Chapter 7. Bibliography

Ahveninen, J., Huang, S., Belliveau, J.W., Chang, W.T. & Hamalainen, M. (2013) Dynamic oscillatory processes governing cued orienting and allocation of auditory attention. *Journal of cognitive neuroscience*, **25**, 1926-1943.

Aitkin, L., Tran, L. & Syka, J. (1994) The responses of neurons in subdivisions of the inferior colliculus of cats to tonal, noise and vocal stimuli. *Experimental brain research*, **98**, 53-64.

Aitkin, L.M., Dickhaus, H., Schult, W. & Zimmermann, M. (1978) External nucleus of inferior colliculus: auditory and spinal somatosensory afferents and their interactions. *Journal of Neurophysiology*, **41**, 837-847.

Akram, S., Presacco, A., Simon, J.Z., Shamma, S.A. & Babadi, B. (2016) Robust decoding of selective auditory attention from MEG in a competing-speaker environment via state-space modeling. *NeuroImage*, **124**, 906-917.

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

Algazi, V.R., Duda, R.O., Thompson, D.M. & Avendano, C. (Year) The CIPIC HRTF Database. IEEE, City. p. 99-102.

Alsius, A., Navarra, J., Campbell, R. & Soto-Faraco, S. (2005) Audiovisual Integration of Speech Falters under High Attention Demands. *Current Biology*, **15**, 839-843.

Alsius, A., Navarra, J. & Soto-Faraco, S. (2007) Attention to touch weakens audiovisual speech integration. *Experimental brain research*, **183**, 399-404.

Anderson, A.J., Kiela, D., Clark, S. & Poesio, M. (2017) Visually Grounded and Textual Semantic Models Differentially Decode Brain Activity Associated with Concrete and Abstract Nouns. *Transactions of the Association for Computational Linguistics*, **5**, 17-30.

Antunes, F.M., Nelken, I., Covey, E. & Malmierca, M.S. (2010) Stimulus-Specific Adaptation in the Auditory Thalamus of the Anesthetized Rat. *PLOS ONE*, **5**, e14071.

Arnal, L.H., Morillon, B., Kell, C.A. & Giraud, A.L. (2009) Dual neural routing of visual facilitation in speech processing. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, **29**, 13445-13453.

Auer Jr, E.T. & Bernstein, L.E. (1997) Speechreading and the structure of the lexicon: Computationally modeling the effects of reduced phonetic distinctiveness on lexical uniqueness. *The Journal of the Acoustical Society of America*, **102**, 3704-3710.

Baart, M., Vroomen, J., Shaw, K. & Bortfeld, H. (2014) Degrading phonetic information affects matching of audiovisual speech in adults, but not in infants. *Cognition*, **130**, 31-43.

Bajo, V.M., Nodal, F.R., Moore, D.R. & King, A.J. (2010) The descending corticocollicular pathway mediates learning-induced auditory plasticity. *Nature neuroscience*, **13**, 253.

Barth, D.S., Goldberg, N., Brett, B. & Di, S. (1995) The spatiotemporal organization of auditory, visual, and auditory-visual evoked potentials in rat cortex. *Brain research*, **678**, 177-190.

Bates, D., Mächler, M., Bolker, B. & Walker, S. (2014) Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Beauchamp, M.S., Lee, K.E., Argall, B.D. & Martin, A. (2004) Integration of auditory and visual information about objects in superior temporal sulcus. *Neuron*, **41**, 809-823.

Bednar, A. & Lalor, E.C. (2018) Neural tracking of auditory motion is reflected by delta phase and alpha power of EEG. *NeuroImage*, **181**, 683-691.

Benjamini, Y. & Hochberg, Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, **57**, 289-300.

Berman, A.L. (1961) Interaction of cortical responses to somatic and auditory stimuli in anterior ectosylvian gyrus of cat. *J Neurophysiol*, **24**, 608-620.

Bernstein, L.E., Auer, E.T., Jr., Moore, J.K., Ponton, C.W., Don, M. & Singh, M. (2002) Visual speech perception without primary auditory cortex activation. *Neuroreport*, **13**, 311-315.

Bernstein, L.E., Auer, E.T., Jr., Wagner, M. & Ponton, C.W. (2008) Spatiotemporal dynamics of audiovisual speech processing. *NeuroImage*, **39**, 423-435.

Bernstein, L.E., Jiang, J., Pantazis, D., Lu, Z.L. & Joshi, A. (2011) Visual phonetic processing localized using speech and nonspeech face gestures in video and point-light displays. *Hum Brain Mapp*, **32**, 1660-1676.

Bernstein, L.E. & Liebenthal, E. (2014) Neural pathways for visual speech perception. *Front Neurosci*, **8**, 386.

Bertelson, P., Vroomen, J., De Gelder, B. & Driver, J. (2000) The ventriloquist effect does not depend on the direction of deliberate visual attention. *Perception & psychophysics*, **62**, 321-332.

Besle, J., Fischer, C., Bidet-Caulet, A., Lecaignard, F., Bertrand, O. & Giard, M.H. (2008) Visual activation and audiovisual interactions in the auditory cortex during speech perception: intracranial recordings in humans. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, **28**, 14301-14310.

Besle, J., Fort, A., Delpuech, C. & Giard, M.H. (2004a) Bimodal speech: early suppressive visual effects in human auditory cortex. *The European journal of neuroscience*, **20**, 2225-2234.

Besle, J., Fort, A. & Giard, M.-H. (2004b) Interest and validity of the additive model in electrophysiological studies of multisensory interactions. *Cognitive Processing*, **5**, 189-192.

Bidelman, G.M., Brown, B., Mankel, K. & Nelms Price, C. (2020) Psychobiological Responses Reveal Audiovisual Noise Differentially Challenges Speech Recognition. *Ear and Hearing*, **41**, 268-277.

Bilenko, N.Y. & Gallant, J.L. (2016) Pyrcca: Regularized Kernel Canonical Correlation Analysis in Python and Its Applications to Neuroimaging. *Frontiers in Neuroinformatics*, **10**.

Binder, J.R., Frost, J.A., Hammeke, T.A., Bellgowan, P.S.F., Springer, J.A., Kaufman, J.N. & Possing, E.T. (2000) Human Temporal Lobe Activation by Speech and Nonspeech Sounds. *Cerebral Cortex*, **10**, 512-528.

Bok, S.T. (1959) Histonomy of the cerebral cortex. *BJS (British Journal of Surgery)*, **47**, 454-454.

Bourguignon, M., Baart, M., Kapnoula, E.C. & Molinaro, N. (2020) Lip-Reading Enables the Brain to Synthesize Auditory Features of Unknown Silent Speech. *The Journal of Neuroscience*, **40**, 1053-1065.

Brodbeck, C., Hong, L.E. & Simon, J.Z. (2018) Rapid Transformation from Auditory to Linguistic Representations of Continuous Speech. *Current Biology*.

Broderick, M.P., Anderson, A.J., Di Liberto, G.M., Crosse, M.J. & Lalor, E.C. (2018) Electrophysiological correlates of semantic dissimilarity reflect the comprehension of natural, narrative speech. *Current Biology*, **28**, 803-809. e803.

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

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.

Callan, D.E., Jones, J.A., Munhall, K., Kroos, C., Callan, A.M. & Vatikiotis-Bateson, E. (2004) Multisensory integration sites identified by perception of spatial wavelet filtered visual speech gesture information. *Journal of cognitive neuroscience*, **16**, 805-816.

Calvert, G.A., Bullmore, E.T., Brammer, M.J., Campbell, R., Williams, S.C., McGuire, P.K., Woodruff, P.W., Iversen, S.D. & David, A.S. (1997) Activation of auditory cortex during silent lipreading. *Science (New York, N.Y.)*, **276**, 593-596.

Calvert, G.A. & Campbell, R. (2003) Reading speech from still and moving faces: the neural substrates of visible speech. *Journal of cognitive neuroscience*, **15**, 57-70.

Calvert, G.A., Campbell, R. & Brammer, M.J. (2000) Evidence from functional magnetic resonance imaging of crossmodal binding in the human heteromodal cortex. *Current Biology*, **10**, 649-657.

Campbell, R. (2008) The processing of audio-visual speech: empirical and neural bases. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **363**, 1001-1010.

Cappelletta, L. & Harte, N. (Year) Phoneme-to-viseme Mapping for Visual Speech Recognition. ICPRAM (2). City. p. 322-329.

Chan, A.M., Dykstra, A.R., Jayaram, V., Leonard, M.K., Travis, K.E., Gygi, B., Baker, J.M., Eskandar, E., Hochberg, L.R., Halgren, E. & Cash, S.S. (2013) Speech-Specific Tuning of Neurons in Human Superior Temporal Gyrus. *Cerebral Cortex*, **24**, 2679-2693.

Chandrasekaran, B., Kraus, N. & Wong, P.C.M. (2012) Human inferior colliculus activity relates to individual differences in spoken language learning. *Journal of Neurophysiology*, **107**, 1325-1336.

Chandrasekaran, C., Trubanova, A., Stillittano, S., Caplier, A. & Ghazanfar, A.A. (2009) The Natural Statistics of Audiovisual Speech. *PLoS Comput Biol*, **5**, e1000436.

Chang, E.F., Rieger, J.W., Johnson, K., Berger, M.S., Barbaro, N.M. & Knight, R.T. (2010) Categorical speech representation in human superior temporal gyrus. *Nature neuroscience*, **13**, 1428.

Cherry, E.C. (1953) Some Experiments on the Recognition of Speech, with One and with Two Ears. *The Journal of the Acoustical Society of America*, **25**, 975-979.

Chomsky, N. & Halle, M. (1968) *The sound pattern of English*. Harper & Row, New York.

Cibelli, E.S., Leonard, M.K., Johnson, K. & Chang, E.F. (2015) The influence of lexical statistics on temporal lobe cortical dynamics during spoken word listening. *Brain and Language*, **147**, 66-75.

Clarey, J.C., Paolini, A.G., Grayden, D.B., Burkitt, A.N. & Clark, G.M. (2004) Ventral cochlear nucleus coding of voice onset time in naturally spoken syllables. *Hearing Research*, **190**, 37-59.

Combrisson, E. & Jerbi, K. (2015) Exceeding chance level by chance: The caveat of theoretical chance levels in brain signal classification and statistical assessment of decoding accuracy. *J Neurosci Methods*, **250**, 126-136.

Correa, N.M., Adali, T., Li, Y. & Calhoun, V.D. (2010) Canonical Correlation Analysis for Data Fusion and Group Inferences. *IEEE Signal Processing Magazine*, **27**, 39-50.

Crosse, M.J., Butler, J.S. & Lalor, E.C. (2015a) Congruent Visual Speech Enhances Cortical Entrainment to Continuous Auditory Speech in Noise-Free Conditions. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, **35**, 14195-14204.

Crosse, M.J., Di Liberto, G.M., Bednar, A. & Lalor, E.C. (2016a) The Multivariate Temporal Response Function (mTRF) Toolbox: A MATLAB Toolbox for Relating Neural Signals to Continuous Stimuli. *Front Hum Neurosci*, **10**, 604.

Crosse, M.J., Di Liberto, G.M. & Lalor, E.C. (2016b) Eye Can Hear Clearly Now: Inverse Effectiveness in Natural Audiovisual Speech Processing Relies on Long-Term Crossmodal Temporal Integration. *The Journal of Neuroscience*, **36**, 9888-9895.

Crosse, M.J., ElShafei, H.A., Foxe, J.J. & Lalor, E.C. (Year) Investigating the temporal dynamics of auditory cortical activation to silent lipreading. 2015 7th International IEEE/EMBS Conference on Neural Engineering (NER). IEEE, City. p. 308-311.

Cruz-Cano, R. & Lee, M.-L.T. (2014) Fast regularized canonical correlation analysis. *Computational Statistics & Data Analysis*, **70**, 88-100.

Das, N., Biesmans, W., Bertrand, A. & Francart, T. (2016) The effect of head-related filtering and ear-specific decoding bias on auditory attention detection. *J Neural Eng*, **13**, 056014.

Daube, C., Ince, R.A.A. & Gross, J. (2019) Simple Acoustic Features Can Explain Phoneme-Based Predictions of Cortical Responses to Speech. *Current Biology*, **29**, 1924-1937.e1929.

Davis, C. & Kim, J. (2006) Audio-visual speech perception off the top of the head. *Cognition*, **100**, B21-B31.

de Cheveigné, A. & Arzounian, D. (2018) Robust detrending, rereferencing, outlier detection, and inpainting for multichannel data. *NeuroImage*, **172**, 903-912.

de Cheveigne, A., Wong, D.D.E., Di Liberto, G.M., Hjortkjaer, J., Slaney, M. & Lalor, E. (2018) Decoding the auditory brain with canonical component analysis. *NeuroImage*, **172**, 206-216.

de Heer, W.A., Huth, A.G., Griffiths, T.L., Gallant, J.L. & Theunissen, F.E. (2017) The hierarchical cortical organization of human speech processing. *Journal of Neuroscience*, **37**, 6539-6557.

Delorme, A. & Makeig, S. (2004) EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, **134**, 9-21.

Denk, F., Grzybowski, M., Ernst, S.M.A., Kollmeier, B., Debener, S. & Bleichner, M.G. (2018) Event-Related Potentials Measured From In and Around the Ear Electrodes Integrated in a Live Hearing Device for Monitoring Sound Perception. *Trends in hearing*, **22**, 2331216518788219.

DeWitt, I. & Rauschecker, J.P. (2012) Phoneme and word recognition in the auditory ventral stream. **109**, E505-E514.

Di Liberto, G.M., O'Sullivan, J.A. & Lalor, E.C. (2015) Low-Frequency Cortical Entrainment to Speech Reflects Phoneme-Level Processing. *Current Biology*, **25**, 2457-2465.

Ding, N., Chatterjee, M. & Simon, J.Z. (2013) Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure. *NeuroImage*, **88c**, 41-46.

Ding, N. & Simon, J.Z. (2012a) Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences*, **109**, 11854-11859.

Ding, N. & Simon, J.Z. (2012b) Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. *Journal of neurophysiology*, **107**, 78-89.

Ding, N. & Simon, J.Z. (2013) Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, **33**, 5728-5735.

Ding, N. & Simon, J.Z. (2014) Cortical Entrainment to Continuous Speech: Functional Roles and Interpretations. *Frontiers in Human Neuroscience*, **8**.

Donhauser, P.W. & Baillet, S. (2020) Two Distinct Neural Timescales for Predictive Speech Processing. *Neuron*, **105**, 385-393.e389.

Driver, J. (1996) Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading. *Nature*, **381**, 66-68.

Ernst, M.O. & Banks, M.S. (2002) Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, **415**, 429-433.

Eskelund, K., Tuomainen, J. & Andersen, T.S. (2011) Multistage audiovisual integration of speech: dissociating identification and detection. *Experimental brain research*, **208**, 447-457.

Fant, G. (1966) A note on vocal tract size factors and non-uniform F-pattern scalings. *Speech Transmission Laboratory Quarterly Progress and Status Report*, **1**, 22-30.

Feng, W., Störmer, V.S., Martinez, A., McDonald, J.J. & Hillyard, S.A. (2017) Involuntary orienting of attention to a sound desynchronizes the occipital alpha rhythm and improves visual perception. *NeuroImage*, **150**, 318-328.

Files, B.T., Auer, E.T. & Bernstein, L.E. (2013) The visual mismatch negativity elicited with visual speech stimuli. *Frontiers in human neuroscience*, **7**, 371.

Fisher, C.G. (1968) Confusions among visually perceived consonants. *Journal of speech and hearing research*, **11**, 796-804.

Formisano, E., De Martino, F., Bonte, M. & Goebel, R. (2008) "Who" Is Saying "What"? Brain-Based Decoding of Human Voice and Speech. *Science (New York, N.Y.)*, **322**, 970-973.

Foxe, J.J., Morocz, I.A., Murray, M.M., Higgins, B.A., Javitt, D.C. & Schroeder, C.E. (2000) Multisensory auditory-somatosensory interactions in early cortical processing revealed by high-density electrical mapping. *Brain research. Cognitive brain research*, **10**, 77-83.

Foxe, J.J., Simpson, G.V. & Ahlfors, S.P. (1998) Parieto-occipital approximately 10 Hz activity reflects anticipatory state of visual attention mechanisms. *Neuroreport*, **9**, 3929-3933.

Foxe, J.J. & Snyder, A.C. (2011) The Role of Alpha-Band Brain Oscillations as a Sensory Suppression Mechanism during Selective Attention. *Frontiers in psychology*, **2**, 154.

French, N.R. & Steinberg, J.C. (1947) Factors governing the intelligibility of speech sounds. *The journal of the Acoustical society of America*, **19**, 90-119.

Frey, J.N., Mainy, N., Lachaux, J.-P., Müller, N., Bertrand, O. & Weisz, N. (2014) Selective Modulation of Auditory Cortical Alpha Activity in an Audiovisual Spatial Attention Task. *The Journal of Neuroscience*, **34**, 6634.

Fu, K.M., Foxe, J.J., Murray, M.M., Higgins, B.A., Javitt, D.C. & Schroeder, C.E. (2001) Attention-dependent suppression of distracter visual input can be cross-modally cued as indexed by anticipatory parieto-occipital alpha-band oscillations. *Brain research. Cognitive brain research*, **12**, 145-152.

Fuglsang, S.A., Dau, T. & Hjortkjær, J. (2017) Noise-robust cortical tracking of attended speech in real-world acoustic scenes. *NeuroImage*, **156**, 435-444.

Gay, T. (1978) Effect of speaking rate on vowel formant movements. *The journal of the Acoustical society of America*, **63**, 223-230.

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**, 473-490.

Golumbic, E.M.Z., Ding, N., Bickel, S., Lakatos, P., Schevon, C.A., McKhann, G.M., Goodman, R.R., Emerson, R., Mehta, A.D. & Simon, J.Z. (2013a) Mechanisms underlying selective neuronal tracking of attended speech at a "cocktail party". *Neuron*, **77**, 980-991.

Golumbic, E.Z., Cogan, G.B., Schroeder, C.E. & Poeppel, D. (2013b) Visual Input Enhances Selective Speech Envelope Tracking in Auditory Cortex at a 'Cocktail Party'. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, **33**, 1417-1426.

Gorman, K., Howell, J. & Wagner, M. (2011) Prosodylab-aligner: A tool for forced alignment of laboratory speech. *2011*, **39**, 2.

Grant, K.W. & Seitz, P.F. (2000) The use of visible speech cues for improving auditory detection of spoken sentences. *J Acoust Soc Am*, **108**, 1197-1208.

Gruters, K.G., Murphy, D.L., Jenson, C.D., Smith, D.W., Shera, C.A. & Groh, J.M. (2018) The eardrums move when the eyes move: A multisensory effect on the mechanics of hearing. *Proceedings of the National Academy of Sciences*, **115**, E1309-E1318.

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, & Psychophysics*, **77**, 1333-1341.

Gwilliams, L., King, J.-R., Marantz, A. & Poeppel, D. (2020) Neural dynamics of phoneme sequencing in real speech jointly encode order and invariant content. *bioRxiv*, 2020.2004.2004.025684.

Hamilton, L.S., Edwards, E. & Chang, E.F. (2018) A Spatial Map of Onset and Sustained Responses to Speech in the Human Superior Temporal Gyrus. *Current Biology*, **28**, 1860-1871.e1864.

Hamilton, L.S. & Huth, A.G. (2018) The revolution will not be controlled: natural stimuli in speech neuroscience. *Language, Cognition and Neuroscience*, 1-10.

Handy, T.C. (2005) *Event-related potentials: A methods handbook*. MIT press.

Hardoon, D.R., Szedmak, S. & Shawe-Taylor, J. (2004) Canonical Correlation Analysis: An Overview with Application to Learning Methods. *Neural Computation*, **16**, 2639-2664.

Haufe, S., Meinecke, F., Gorgen, K., Dahne, S., Haynes, J.D., Blankertz, B. & Biessmann, F. (2014) On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, **87**, 96-110.

Hauswald, A., Lithari, C., Collignon, O., Leonardelli, E. & Weisz, N. (2018) A Visual Cortical Network for Deriving Phonological Information from Intelligible Lip Movements. *Current Biology*, **28**, 1453-1459.e1453.

Hickok, G. & Poeppel, D. (2007) The cortical organization of speech processing. *Nature Reviews Neuroscience*, **8**, 393-402.

Holler, J. & Levinson, S.C. (2019) Multimodal language processing in human communication. *Trends in cognitive sciences*, **23**, 639-652.

Holt, L.L. & Lotto, A.J. (2010) Speech perception as categorization. *Attention, perception & psychophysics*, **72**, 1218-1227.

Hotelling, H. (1936) RELATIONS BETWEEN TWO SETS OF VARIATES. *Biometrika*, **28**, 321-377.

Houtgast, T. & Steeneken, H.J. (1973) The modulation transfer function in room acoustics as a predictor of speech intelligibility. *Acta Acustica United With Acustica*, **28**, 66-73.

Howard III, M.A., Volkov, I.O., Abbas, P.J., Damasio, H., Ollendieck, M.C. & Granner, M.A. (1996) A chronic microelectrode investigation of the tonotopic organization of human auditory cortex. *Brain research*, **724**, 260-264.

Huth, A.G., de Heer, W.A., Griffiths, T.L., Theunissen, F.E. & Gallant, J.L. (2016) Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, **532**, 453-458.

Irino, T. & Patterson, R.D. (2006) A Dynamic Compressive Gammachirp Auditory Filterbank. *IEEE transactions on audio, speech, and language processing*, **14**, 2222-2232.

Jääskeläinen, I.P., Ojanen, V., Ahveninen, J., Auranen, T., Levänen, S., Möttönen, R., Tarnanen, I. & Sams, M. (2004) Adaptation of neuromagnetic N1 responses to phonetic stimuli by visual speech in humans. *Neuroreport*, **15**, 2741-2744.

Jakobson, R., Fant, C.G. & Halle, M. (1951) Preliminaries to speech analysis: The distinctive features and their correlates.

Jiang, J., Alwan, A., Keating, P., Auer, E., Jr. & Bernstein, L. (2002) On the Relationship between Face Movements, Tongue Movements, and Speech Acoustics. *EURASIP J. Adv. Signal Process.*, **2002**, 1174–1188.

Kanwal, J.S. & Rauschecker, J.P. (2007) Auditory cortex of bats and primates: managing species-specific calls for social communication. *Frontiers in bioscience: a journal and virtual library*, **12**, 4621.

Karas, P.J., Magnotti, J.F., Metzger, B.A., Zhu, L.L., Smith, K.B., Yoshor, D. & Beauchamp, M.S. (2019) The visual speech head start improves perception and reduces superior temporal cortex responses to auditory speech. *eLife*, **8**, e48116.

Kayser, C. & Logothetis, N.K. (2009) Directed interactions between auditory and superior temporal cortices and their role in sensory integration. *Frontiers in integrative neuroscience*, **3**, 7.

Kayser, C., Petkov, C.I. & Logothetis, N.K. (2008) Visual Modulation of Neurons in Auditory Cortex. *Cerebral Cortex*, **18**, 1560-1574.

Kelly, S.P., Lalor, E.C., Reilly, R.B. & Foxe, J.J. (2006) Increases in alpha oscillatory power reflect an active retinotopic mechanism for distracter suppression during sustained visuospatial attention. *J Neurophysiol*, **95**, 3844-3851.

Kent, R.D. & Minifie, F.D. (1977) Coarticulation in recent speech production models. *Journal of Phonetics*, **5**, 115-133.

Kerlin, J.R., Shahin, A.J. & Miller, L.M. (2010) Attentional gain control of ongoing cortical speech representations in a "cocktail party". *The Journal of neuroscience : the official journal of the Society for Neuroscience*, **30**, 620-628.

Khalighinejad, B., Cruzatto da Silva, G. & Mesgarani, N. (2017) Dynamic Encoding of Acoustic Features in Neural Responses to Continuous Speech. *The Journal of Neuroscience*, **37**, 2176-2185.

King, J.-R., Charton, F., Lopez-Paz, D. & Oquab, M. (2020) Back-to-back regression: Disentangling the influence of correlated factors from multivariate observations. *NeuroImage*, **220**, 117028.

Klatt, D.H. (1979) Speech perception: A model of acoustic–phonetic analysis and lexical access. *Journal of phonetics*, **7**, 279-312.

Klatt, D.H. & Klatt, L.C. (1990) Analysis, synthesis, and perception of voice quality variations among female and male talkers. *the Journal of the Acoustical Society of America*, **87**, 820-857.

Klimesch, W. (2012) Alpha-band oscillations, attention, and controlled access to stored information. *Trends in cognitive sciences*, **16**, 606-617.

Klucharev, V., Mottonen, R. & Sams, M. (2003) Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audiovisual speech perception. *Brain research. Cognitive brain research*, **18**, 65-75.

Körding, K.P. & Wolpert, D.M. (2004) Bayesian integration in sensorimotor learning. *Nature*, **427**, 244-247.

Kuperberg, G.R. & Jaeger, T.F. (2016) What do we mean by prediction in language comprehension? *Language, Cognition and Neuroscience*, **31**, 32-59.

Lalor, E.C. & Foxe, J.J. (2010) Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. *The European journal of neuroscience*, **31**, 189-193.

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

Lee, Y.-S., Turkeltaub, P., Granger, R. & Raizada, R.D.S. (2012) Categorical Speech Processing in Broca's Area: An fMRI Study Using Multivariate Pattern-Based Analysis. *The Journal of Neuroscience*, **32**, 3942-3948.

Lemus, L., Hernandez, A., Luna, R., Zainos, A. & Romo, R. (2010) Do sensory cortices process more than one sensory modality during perceptual judgments? *Neuron*, **67**, 335-348.

Leonard, M.K., Baud, M.O., Sjerps, M.J. & Chang, E.F. (2016) Perceptual restoration of masked speech in human cortex. *Nature communications*, **7**, 1-9.

Leonard, M.K., Bouchard, K.E., Tang, C. & Chang, E.F. (2015) Dynamic Encoding of Speech Sequence Probability in Human Temporal Cortex. *The Journal of Neuroscience*, **35**, 7203-7214.

Lettieri, G., Handjaras, G., Ricciardi, E., Leo, A., Papale, P., Betta, M., Pietrini, P. & Cecchetti, L. (2019) Emotionotopy in the human right temporo-parietal cortex. *Nature Communications*, **10**, 5568.

Leurgans, S.E., Moyeed, R.A. & Silverman, B.W. (1993) Canonical Correlation Analysis When the Data are Curves. *Journal of the Royal Statistical Society: Series B (Methodological)*, **55**, 725-740.

Liberman, A.M., Cooper, F.S., Shankweiler, D.P. & Studdert-Kennedy, M. (1967) Perception of the speech code. *Psychological review*, **74**, 431-461.

Liebenthal, E., Binder, J.R., Spitzer, S.M., Possing, E.T. & Medler, D.A. (2005) Neural Substrates of Phonemic Perception. *Cerebral Cortex*, **15**, 1621-1631.

Liebenthal, E., Desai, R., Ellingson, M.M., Ramachandran, B., Desai, A. & Binder, J.R. (2010) Specialization along the left superior temporal sulcus for auditory categorization. *Cerebral cortex (New York, N.Y. : 1991)*, **20**, 2958-2970.

Ludman, C.N., Summerfield, A.Q., Hall, D., Elliott, M., Foster, J., Hykin, J.L., Bowtell, R. & Morris, P.G. (2000) Lip-reading ability and patterns of cortical activation studied using fMRI. *British journal of audiology*, **34**, 225-230.

Luo, H., Liu, Z. & Poeppel, D. (2010) Auditory cortex tracks both auditory and visual stimulus dynamics using low-frequency neuronal phase modulation. *PLoS biology*, **8**, e1000445.

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**, e4638.

MacDonald, J., Andersen, S. & Bachmann, T. (2000) Hearing by Eye: How Much Spatial Degradation can Be Tolerated? *Perception*, **29**, 1155-1168.

Mazaheri, A., van Schouwenburg, M.R., Dimitrijevic, A., Denys, D., Cools, R. & Jensen, O. (2014) Region-specific modulations in oscillatory alpha activity serve to facilitate processing in the visual and auditory modalities. *NeuroImage*, **87**, 356-362.

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**, 762-776.

McGurk, H. & Macdonald, J. (1976) Hearing lips and seeing voices. *Nature*, **264**, 746-748.

Meredith, M.A. & Stein, B.E. (1983) Interactions among converging sensory inputs in the superior colliculus. *Science (New York, N.Y.)*, **221**, 389-391.

Meredith, M.A. & Stein, B.E. (1986) Spatial factors determine the activity of multisensory neurons in cat superior colliculus. *Brain research*, **365**, 350-354.

Meredith, M.A. & Stein, B.E. (1993) *The merging of the senses*. MIT Press, United States of America.

Mesgarani, N. & Chang, E.F. (2012) Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, **485**, 233.

Mesgarani, N., Cheung, C., Johnson, K. & Chang, E.F. (2014) Phonetic Feature Encoding in Human Superior Temporal Gyrus. *Science (New York, N.Y.)*, **343**, 1006-1010.

Mesgarani, N., David, S.V., Fritz, J.B. & Shamma, S.A. (2008) Phoneme representation and classification in primary auditory cortex. *The Journal of the Acoustical Society of America*, **123**, 899-909.

Mesgarani, N., David, S.V., Fritz, J.B. & Shamma, S.A. (2009) Influence of context and behavior on stimulus reconstruction from neural activity in primary auditory cortex. *Journal of neurophysiology*, **102**, 3329-3339.

Metzger, B.A., Magnotti, J.S., Wang, Z., Nesbitt, E., Karas, P.J., Yoshor, D. & Beauchamp, M.S. (2020) Responses to Visual Speech in Human Posterior Superior Temporal Gyrus Examined with iEEG Deconvolution. *bioRxiv*, 2020.2004.2016.045716.

Miller, J.L. & Baer, T. (1983) Some effects of speaking rate on the production of/b/and/w. *The Journal of the Acoustical Society of America*, **73**, 1751-1755.

Miller, L.M. & D'Esposito, M. (2005) Perceptual fusion and stimulus coincidence in the cross-modal integration of speech. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, **25**, 5884-5893.

Miran, S., Akram, S., Sheikhattar, A., Simon, J.Z., Zhang, T. & Babadi, B. (2018) Real-Time Tracking of Selective Auditory Attention From M/EEG: A Bayesian Filtering Approach. *Frontiers in Neuroscience*, **12**.

Mirkovic, B., Debener, S., Jaeger, M. & Vos, M.D. (2015) Decoding the attended speech stream with multi-channel EEG: implications for online, daily-life applications. *Journal of Neural Engineering*, **12**, 046007.

Mitchell, T.M., Shinkareva, S.V., Carlson, A., Chang, K.-M., Malave, V.L., Mason, R.A. & Just, M.A. (2008) Predicting Human Brain Activity Associated with the Meanings of Nouns. *Science (New York, N.Y.)*, **320**, 1191-1195.

Molholm, S., Ritter, W., Murray, M.M., Javitt, D.C., Schroeder, C.E. & Foxe, J.J. (2002) Multisensory auditory–visual interactions during early sensory processing in humans: a high-density electrical mapping study. *Cognitive Brain Research*, **14**, 115-128.

Morís Fernández, L., Visser, M., Ventura-Campos, N., Ávila, C. & Soto-Faraco, S. (2015) Top-down attention regulates the neural expression of audiovisual integration. *NeuroImage*, **119**, 272-285.

Möttönen, R., Krause, C.M., Tiippana, K. & Sams, M. (2002) Processing of changes in visual speech in the human auditory cortex. *Brain research. Cognitive brain research*, **13**, 417-425.

Möttönen, R., Schürmann, M. & Sams, M. (2004) Time course of multisensory interactions during audiovisual speech perception in humans: a magnetoencephalographic study. *Neuroscience letters*, **363**, 112-115.

Munhall, K.G., Jones, J.A., Callan, D.E., Kuratate, T. & Vatikiotis-Bateson, E. (2004) Visual Prosody and Speech Intelligibility: Head Movement Improves Auditory Speech Perception. *Psychological Science*, **15**, 133-137.

Musacchia, G., Sams, M., Nicol, T. & Kraus, N. (2006) Seeing speech affects acoustic information processing in the human brainstem. *Experimental brain research*, **168**, 1-10.

Nath, A.R. & Beauchamp, M.S. (2011) Dynamic changes in superior temporal sulcus connectivity during perception of noisy audiovisual speech. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, **31**, 1704-1714.

Neti, C., Potamianos, G., Luettin, J., Matthews, I., Glotin, H., Vergyri, D., Sison, J. & Mashari, A. (2000) Audio visual speech recognition. IDIAP.

Nourski, K.V. (2017) Auditory processing in the human cortex: An intracranial electrophysiology perspective. *Laryngoscope investigative otolaryngology*, **2**, 147-156.

Nourski, K.V., Steinschneider, M., McMurray, B., Kovach, C.K., Oya, H., Kawasaki, H. & Howard III, M.A. (2014) Functional organization of human auditory cortex: investigation of response latencies through direct recordings. *NeuroImage*, **101**, 598-609.

O'Sullivan, A.E., Crosse, M.J., Di Liberto, G.M. & Lalor, E.C. (2017) Visual Cortical Entrainment to Motion and Categorical Speech Features during Silent Lipreading. *Frontiers in human neuroscience*, **10**, 679.

O'Sullivan, J.A., Power, A.J., Mesgarani, N., Rajaram, S., Foxe, J.J., Shinn-Cunningham, B.G., Slaney, M., Shamma, S.A. & Lalor, E.C. (2014) Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG. *Cerebral Cortex*.

O'Sullivan, J., Chen, Z., Herrero, J., McKhann, G., M., Sheth, S., A. , Mehta, A., D. & Mesgarani, N. (2017) Neural decoding of attentional selection in multi-speaker environments without access to clean sources. *Journal of Neural Engineering*, **14**, 056001.

Obleser, J. & Kayser, C. (2019) Neural Entrainment and Attentional Selection in the Listening Brain. *Trends in cognitive sciences*, **23**, 913-926.

Öhman, S.E. (1966) Coarticulation in VCV utterances: Spectrographic measurements. *The Journal of the Acoustical Society of America*, **39**, 151-168.

Okada, K., Venezia, J.H., Matchin, W., Saberi, K. & Hickok, G. (2013) An fMRI study of audiovisual speech perception reveals multisensory interactions in auditory cortex. *PloS one*, **8**, e68959.

Oliver, D.L. & Morest, D.K. (1984) The central nucleus of the inferior colliculus in the cat. *J Comp Neurol*, **222**, 237-264.

Overath, T., McDermott, J.H., Zarate, J.M. & Poeppel, D. (2015) The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. *Nature neuroscience*, **18**, 903-911.

Ozker, M., Yoshor, D. & Beauchamp, M.S. (2018) Frontal cortex selects representations of the talker's mouth to aid in speech perception. *eLife*, **7**, e30387.

Pannese, A., Grandjean, D. & Frühholz, S. (2015) Subcortical processing in auditory communication. *Hearing Research*, **328**, 67-77.

Park, H., Kayser, C., Thut, G. & Gross, J. (2016) Lip movements entrain the observers' low-frequency brain oscillations to facilitate speech intelligibility. *eLife*, **5**, e14521.

Parsons, T.W. (1987) *Voice and speech processing*. New York: McGraw-Hill College.

Peelle, J.E. & Sommers, M.S. (2015) Prediction and constraint in audiovisual speech perception. *Cortex; a journal devoted to the study of the nervous system and behavior*, **68**, 169-181.

Pekkola, J., Ojanen, V., Autti, T., Jaaskelainen, I.P., Mottonen, R., Tarkiainen, A. & Sams, M. (2005) Primary auditory cortex activation by visual speech: an fMRI study at 3 T. *Neuroreport*, **16**, 125-128.

Pereira, F., Lou, B., Pritchett, B., Ritter, S., Gershman, S.J., Kanwisher, N., Botvinick, M. & Fedorenko, E. (2018) Toward a universal decoder of linguistic meaning from brain activation. *Nat Commun*, **9**, 963.

Pfurtscheller, G. & Klimesch, W. (1991) Event-related desynchronization during motor behavior and visual information processing. *Electroencephalography and clinical neurophysiology. Supplement*, **42**, 58-65.

Plass, J., Brang, D., Suzuki, S. & Grabowecky, M. (2019) Vision Perceptually Restores Auditory Spectral Dynamics in Speech. *PsyArXiv*.

Purves, D., Augustine, G.J., Fitzpatrick, D., Hall, W., LaMantia, A., McNamara, J. & White, L. (2008) Neuroscience. *Sunderland, MA, USA: Sinauer Associates, Inc.*

Puvvada, K.C. & Simon, J.Z. (2017) Cortical Representations of Speech in a Multi-talker Auditory Scene. *The Journal of Neuroscience*.

Rampinini, A.C., Handjaras, G., Leo, A., Cecchetti, L., Betta, M., Marotta, G., Ricciardi, E. & Pietrini, P. (2019) Formant Space Reconstruction From Brain Activity in Frontal and Temporal Regions Coding for Heard Vowels. *Frontiers in Human Neuroscience*, **13**.

Rasmussen, G.L. (1964) Anatomic relationships of the ascending and descending auditory systems. *Neurological aspects of auditory and vestibular disorders*, **1**, 5-19.

Rauschecker, J.P. & Scott, S.K. (2009) Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. *Nature neuroscience*, **12**, 718-724.

Reisberg, D., Mclean, J. & Goldfield, A. (1987) Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli.

Romei, V., Brodbeck, V., Michel, C., Amedi, A., Pascual-Leone, A. & Thut, G. (2008) Spontaneous Fluctuations in Posterior α-Band EEG Activity Reflect Variability in Excitability of Human Visual Areas. *Cerebral Cortex*, **18**, 2010-2018.

Rosenblum, L.D., Johnson, J.A. & Saldaña, H.M. (1996) Point-light facial displays enhance comprehension of speech in noise. *Journal of speech and hearing research*, **39**, 1159-1170.

Ross, L.A., Saint-Amour, D., Leavitt, V.M., Javitt, D.C. & Foxe, J.J. (2007) Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environments. *Cerebral cortex (New York, N.Y. : 1991)*, **17**, 1147-1153.

Saint-Amour, D., De Sanctis, P., Molholm, S., Ritter, W. & Foxe, J.J. (2007) Seeing voices: High-density electrical mapping and source-analysis of the multisensory mismatch negativity evoked during the McGurk illusion. *Neuropsychologia*, **45**, 587-597.

Sams, M., Aulanko, R., Hamalainen, M., Hari, R., Lounasmaa, O.V., Lu, S.T. & Simola, J. (1991) Seeing speech: visual information from lip movements modifies activity in the human auditory cortex. *Neuroscience letters*, **127**, 141-145.

Samuel, A.G. (2011) Speech perception. *Annual review of psychology*, **62**, 49-72.

Schepers, I.M., Yoshor, D. & Beauchamp, M.S. (2015) Electrocorticography Reveals Enhanced Visual Cortex Responses to Visual Speech. *Cerebral Cortex*, **25**, 4103-4110.

Schreiner, C.E. & Winer, J.A. (2005) *The inferior colliculus*. Springer.

Schroeder, C.E. & Foxe, J. (2005) Multisensory contributions to low-level, 'unisensory' processing. *Current Opinion in Neurobiology*, **15**, 454-458.

Schroeder, C.E., Lakatos, P., Kajikawa, Y., Partan, S. & Puce, A. (2008) Neuronal oscillations and visual amplification of speech. *Trends in cognitive sciences*, **12**, 106-113.

Schwartz, J.-L. & Savariaux, C. (2014) No, There Is No 150 ms Lead of Visual Speech on Auditory Speech, but a Range of Audiovisual Asynchronies Varying from Small Audio Lead to Large Audio Lag. *PLOS Computational Biology*, **10**, e1003743.

Schwartz, J.L., Berthommier, F. & Savariaux, C. (2004) Seeing to hear better: evidence for early audio-visual interactions in speech identification. *Cognition*, **93**, B69-78.

Scott, S.K. & Johnsrude, I.S. (2003) The neuroanatomical and functional organization of speech perception. *Trends Neurosci*, **26**, 100-107.

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

Seltzer, B. & Pandya, D.N. (1978) Afferent cortical connections and architectonics of the superior temporal sulcus and surrounding cortex in the rhesus monkey. *Brain research*, **149**, 1-24.

Shahin, A.J., Backer, K.C., Rosenblum, L.D. & Kerlin, J.R. (2018) Neural mechanisms underlying cross-modal phonetic encoding. *Journal of Neuroscience*, **38**, 1835-1849.

Shatzer, H., Shen, S., Kerlin, J.R., Pitt, M.A. & Shahin, A.J. (2018) Neurophysiology underlying influence of stimulus reliability on audiovisual integration. *European Journal of Neuroscience*, **48**, 2836-2848.

Stebbings, K.A., Lesicko, A.M.H. & Llano, D.A. (2014) The auditory corticocollicular system: Molecular and circuit-level considerations. *Hearing Research*, **314**, 51-59.

Stein, B.E. & Meredith, M.A. (1993) *The merging of the senses*. The MIT Press.

Stekelenburg, J.J. & Vroomen, J. (2007) Neural correlates of multisensory integration of ecologically valid audiovisual events. *Journal of cognitive neuroscience*, **19**, 1964-1973.

Stevens, K.N. & House, A.S. (1955) Development of a Quantitative Description of Vowel Articulation. *The Journal of the Acoustical Society of America*, **27**, 484-493.

Stevenson, R.A., Ghose, D., Fister, J.K., Sarko, D.K., Altieri, N.A., Nidiffer, A.R., Kurela, L.R., Siemann, J.K., James, T.W. & Wallace, M.T. (2014) Identifying and quantifying multisensory integration: a tutorial review. *Brain topography*, **27**, 707-730.

Stevenson, R.A. & James, T.W. (2009) Audiovisual integration in human superior temporal sulcus: Inverse effectiveness and the neural processing of speech and object recognition. *NeuroImage*, **44**, 1210-1223.

Strand, J.F., Brown, V.A. & Barbour, D.L. (2019) Talking points: A modulating circle reduces listening effort without improving speech recognition. *Psychonomic Bulletin & Review*, **26**, 291-297.

Sumby, W.H. & Pollack, I. (1954) Visual Contribution to Speech Intelligibility in Noise. *The Journal of the Acoustical Society of America*, **26**, 212-215.

Summerfield, Q. (1979) Use of Visual Information for Phonetic Perception. *Phonetica*, **36**, 314-331.

Summerfield, Q. (1987) Some preliminaries to a comprehensive account of audio-visual speech perception. In Dodd (ed) *Hearing by eye: The psychology of lip-reading*. Lawrence Erlbaum Associates.

Summerfield, Q. (1992) Lipreading and audio-visual speech perception. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, **335**, 71-78.

Talsma, D., Doty, T.J. & Woldorff, M.G. (2007) Selective attention and audiovisual integration: is attending to both modalities a prerequisite for early integration? *Cerebral cortex (New York, N.Y. : 1991)*, **17**, 679-690.

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

Talsma, D. & Woldorff, M.G. (2005) Selective attention and multisensory integration: multiple phases of effects on the evoked brain activity. *Journal of cognitive neuroscience*, **17**, 1098-1114.

ten Oever, S., Schroeder, C.E., Poeppel, D., van Atteveldt, N. & Zion-Golumbic, E. (2014) Rhythmicity and cross-modal temporal cues facilitate detection. *Neuropsychologia*, **63**, 43-50.

Theunissen, F.E., Sen, K. & Doupe, A.J. (2000) Spectral-Temporal Receptive Fields of Nonlinear Auditory Neurons Obtained Using Natural Sounds. *The Journal of Neuroscience*, **20**, 2315-2331.

Thomas, S.M. & Jordan, T.R. (2004) Contributions of oral and extraoral facial movement to visual and audiovisual speech perception. *Journal of experimental psychology. Human perception and performance*, **30**, 873-888.

Tikhonov, A.N. & Arsenin, V.Y. (1977) Solutions of ill-posed problems: Vh Winston.

Tye-Murray, N., Sommers, M. & Spehar, B. (2007) Auditory and visual lexical neighborhoods in audiovisual speech perception. *Trends in amplification*, **11**, 233-241.

Tye-Murray, N., Spehar, B., Myerson, J., Sommers, M.S. & Hale, S. (2011) Crossmodal enhancement of speech detection in young and older adults: Does signal content matter? *Ear Hear.*, **32**, 650.

van Atteveldt, N., Formisano, E., Goebel, R. & Blomert, L. (2004) Integration of Letters and Speech Sounds in the Human Brain. *Neuron*, **43**, 271-282.

Van der Burg, E., Olivers, C.N., Bronkhorst, A.W. & Theeuwes, J. (2008) Pip and pop: nonspatial auditory signals improve spatial visual search. *Journal of experimental psychology. Human perception and performance*, **34**, 1053-1065.

Van der Burg, E., Olivers, C.N., Bronkhorst, A.W. & Theeuwes, J. (2009) Poke and pop: tactile-visual synchrony increases visual saliency. *Neuroscience letters*, **450**, 60-64.

van Wassenhove, V., Grant, K.W. & Poeppel, D. (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.

Varoquaux, G., Sadaghiani, S., Pinel, P., Kleinschmidt, A., Poline, J.B. & Thirion, B. (2010) A group model for stable multi-subject ICA on fMRI datasets. *NeuroImage*, **51**, 288-299.

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

Vatikiotis-Bateson, E. & Munhall, K.G. (2015) Auditory-Visual Speech Processing: Something Doesn't Add Up. *The Handbook of Speech Production*, 178-199.

Vinod, H.D. (1976) Canonical ridge and econometrics of joint production. *Journal of Econometrics*, **4**, 147-166.

von Kriegstein, K., Patterson, R.D. & Griffiths, T.D. (2008) Task-Dependent Modulation of Medial Geniculate Body Is Behaviorally Relevant for Speech Recognition. *Current Biology*, **18**, 1855-1859.

Walden, B.E., Prosek, R.A., Montgomery, A.A., Scherr, C.K. & Jones, C.J. (1977) Effects of Training on the Visual Recognition of Consonants. *Journal of speech and hearing research*, **20**, 130-145.

Wallace, M.N., Anderson, L.A. & Palmer, A.R. (2007) Phase-Locked Responses to Pure Tones in the Auditory Thalamus. *Journal of Neurophysiology*, **98**, 1941-1952.

Weinholtz, C. & Dias, J.W. (2016) Categorical perception of visual speech information. *The Journal of the Acoustical Society of America*, **139**, 2018-2018.

Werner, S. & Noppeney, U. (2010) Superadditive responses in superior temporal sulcus predict audiovisual benefits in object categorization. *Cerebral cortex (New York, N.Y. : 1991)*, **20**, 1829-1842.

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**, 601-615.

Winer, J.A., Larue, D.T., Diehl, J.J. & Hefti, B.J. (1998) Auditory cortical projections to the cat inferior colliculus. *Journal of Comparative Neurology*, **400**, 147-174.

Winter, B. (2013) Linear models and linear mixed effects models in R: Tutorial 11. *arXiv preprint arXiv:1308.5499*.

Woodward, M.F. & Barber, C.G. (1960) Phoneme perception in lipreading. *Journal of speech and hearing research*, **3**, 212-222.

Worden, M.S., Foxe, J.J., Wang, N. & Simpson, G.V. (2000) Anticipatory biasing of visuospatial attention indexed by retinotopically specific alpha-band electroencephalography increases over occipital cortex. *The Journal of neuroscience : the official journal of the Society for Neuroscience*, **20**, Rc63.

Wöstmann, M., Herrmann, B., Maess, B. & Obleser, J. (2016) Spatiotemporal dynamics of auditory attention synchronize with speech. *Proceedings of the National Academy of Sciences*, **113**, 3873-3878.

Wright, T.M., Pelphrey, K.A., Allison, T., McKeown, M.J. & McCarthy, G. (2003) Polysensory interactions along lateral temporal regions evoked by audiovisual speech. *Cerebral cortex (New York, N.Y. : 1991)*, **13**, 1034-1043.

Yang, X., Wang, K. & Shamma, S.A. (1992) Auditory representations of acoustic signals. *IEEE transactions on information theory*, **38**, 824-839.

Yehia, H., Rubin, P. & Vatikiotis-Bateson, E. (1998) Quantitative association of vocal-tract and facial behavior. *Speech Communication*, **26**, 23-43.

Yehia, H.C., Kuratate, T. & Vatikiotis-Bateson, E. (2002) Linking facial animation, head motion and speech acoustics. *Journal of Phonetics*, **30**, 555-568.

Yi, H.G., Leonard, M.K. & Chang, E.F. (2019) The Encoding of Speech Sounds in the Superior Temporal Gyrus. *Neuron*, **102**, 1096-1110.

Yuan, J. & Liberman, M. (2008) *Speaker identification on the SCOTUS corpus*.

Zhu, L.L. & Beauchamp, M.S. (2017) Mouth and Voice: A Relationship between Visual and Auditory Preference in the Human Superior Temporal Sulcus. *J. Neurosci.*, **37**, 2697-2708.