Towards a Greater Understanding of the Neurophysiology of Natural Audiovisual Speech Processing: a System Identification Approach to EEG

by

Michael J. Crosse, B.E., M.Sc.

A dissertation submitted to the

University of Dublin, Trinity College Dublin

for the degree of

Doctor of Philosophy

in the

Department of Electronic and Electrical Engineering

Trinity College Dublin

Dublin, Ireland.

June 2016

Declaration

I declare 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.

Signed,

Michael Crosse

Michael J. Crosse
June 2016

Summary

Seeing a speaker's face as he or she talks can greatly help in understanding what the speaker is saying, especially in adverse hearing conditions – a principle known as inverse effectiveness. This is because the speaker's facial movements not only relay information about *what* the speaker is saying, but also, importantly, *when* the speaker is saying it. Studying how the brain exploits this timing relationship to combine information from continuous auditory and visual speech has traditionally been difficult due to methodological constraints. In contrast, when incongruent auditory and visual information are presented concurrently, it can not only hinder a listener's perception, but even cause him or her to perceive illusory information that was not presented through either modality. Efforts to determine the neurophysiological underpinnings of this phenomenon have also been hampered by out-dated methodological approaches, as well as inaccessibility to state-of-the-art modelling techniques. Here, we introduce a new system identification (SI) framework for investigating these everyday neural processes using relatively inexpensive and non-invasive scalp recordings.

Chapter 3 begins by describing the application SI techniques for studying sensory processing in humans using naturalistic stimuli, specifically in the context of neurophysiology. The aim of this chapter is to introduce a new MATLAB-based SI toolbox, called mTRF Toolbox, developed as part of this research work. Concrete examples demonstrating how to model the relationship between continuous speech stimuli and continuous EEG responses are worked through in full. Several key features of the toolbox are explored and compared to traditional methods and its applications and limitations are discussed.

Chapter 4 examines the role of temporal and contextual congruency in audiovisual (AV) speech processing using the mTRF Toolbox. The development of a novel framework for studying multisensory integration is described, yielding new insights into AV speech processing. Specifically, we show that cortical activity tracks the acoustic speech signal more reliably during congruent AV presentation, while incongruent AV stimuli actually inhibit neural entrainment to speech. The enhancement effect produced by congruent AV stimuli is shown to be most prominent at the rate of syllabic information (2–6 Hz). Furthermore, we demonstrate that neural entrainment to auditory speech during silent lipreading is highly predictive of speech-reading accuracy.

Chapter 5 examines AV speech processing at an acoustic signal-to-noise ratio that maximizes the perceptual benefit conferred by multisensory processing relative to unisensory processing. Here we show that the influence of visual input on the neural tracking of acoustic speech is significantly greater in noisy than in quiet listening conditions, in line with the principle of inverse effectiveness. While envelope tracking during audio-only speech is shown to be greatly reduced by background noise at an early processing stage, it is markedly restored by the addition of visual speech input. We also find that multisensory integration occurs at much lower frequencies in background noise and is predictive of the multisensory gain in behavioural performance at a time lag of ~250 ms. Critically, we demonstrate that inverse effectiveness in natural audiovisual speech processing relies on crossmodal integration over long temporal windows.

Chapter 6 investigates the temporal dynamics of auditory cortical activation associated with silent lipreading by looking at the impact of speech-reading accuracy on neural entrainment to the absent acoustic signal. Specifically, this study provides moderate evidence to suggest that cortical activity in auditory regions is modulated in a way that reflects the temporal dynamics of the absent acoustic information, as if synthesising auditory processing by exploiting correlated visual speech input.

While the non-invasive brain imaging technique implemented in this body of research lacks the spatial resolution to definitively elucidate certain aspects of the neural mechanisms that underpin AV speech processing, its high temporal resolution allows for accurate characterisation of the spectrotemporal dynamics of multisensory integration. The findings presented here provide new, valuable insights into this aspect of AV speech processing in the human brain. The application of this novel SI framework for studying AV speech processing is also considered in the context of clinical disorders with impaired multisensory processing and brain-computer interface technology.

Acknowledgements

This thesis would not have been possible without the invaluable contributions of many people. Therefore, I would to thank:

Ed, for all the time and support he has given me, and for his encouragement and honesty. Thanks for all the memorable trips to SfN and Telluride, and for all the one-sided games of pool!

Richard, for putting me in touch with Ed – this may never have happened if not for that! Thanks for all the references and advice you've given me over the years.

John (Samuel Levingston), for his help with all things multisensory and, most of all, for his constant humour in the lab.

Jim, a.k.a. Big Jim Sullivan, for nine years of friendship and doing science together. Here's to more good times in New York. For science!

Giovanni, for all the enjoyable hours spent exploring new ideas in the lab, and for your friendship. I look forward to more of the same. Thanks for your bed too!

The Neural Engineering crew: Adam, Aisling, Alejandro, Brendan, Céline, Ciara, Denis, Emily, Ger, Hanni, Hugh, Isabelle, Jeremy, Martin, Niamh, Saskia, Terence, Villiam and everyone else past and present. The constant craic in the lab made those long hours very memorable.

Everyone in TCBE, TCIN and GREP (not to mention those who played tag rugby), for making the PhD a great experience and for your friendships.

June and Melanie, for all their help over the four years and for organising such great lab nights out.

Fiona, Hesham, Sean and Alan, for their invaluable assistance with data collection that contributed to the studies presented in this thesis. Also, thanks to everybody who generously volunteered their time to participate in these studies.

Shane Hunt, for his assistance with, what turned out to be, critical experimental hardware, and for all his help with the Analogue labs.

My folks, Paddy and Catherine, and the rest of my family and friends, for their continued support at all times and for helping me see the bigger picture.

Finally, my wonderful girlfriend Dee, for all your hours of proof reading and valuable suggestions. Most of all, for your love and support at all times. Thank you!

This research was kindly supported by the Programme for Research in Third-Level Institutions and co-funded under the European Regional Development fund.

MICHAEL J. CROSSE

Trinity College Dublin
June 2016

Publications Arising from this Thesis

Chapter 3

- **Crosse MJ**, Di Liberto GM, Lalor EC (*in prep*). The Multivariate Temporal Response Function (mTRF) Toolbox: a MATLAB Toolbox for Relating Neural Signals to Continuous Stimuli, for submission to the *Journal of Neuroscience Methods*.

Chapter 4

- **Crosse MJ**, Butler JS, Lalor EC (2015). Congruent Visual Speech Enhances Cortical Entrainment to Continuous Auditory Speech in Noise-Free Conditions. *The Journal of Neuroscience*, 35(42):14195–1420.
- **Crosse MJ**, O'Sullivan JA, Power AJ, Lalor EC (2013). The Effects of Attention and Visual Input on the Representation of Natural Speech in EEG. *35th Annual International Conference of the IEEE/EMBS*, 2800–2803.

Chapter 5

- **Crosse MJ**, Di Liberto GM, Lalor EC (2016). Eye Can Hear Clearly Now: Inverse Effectiveness in Natural Audiovisual Speech Processing Relies on Long-Term Crossmodal Temporal Integration. *The Journal of Neuroscience*, accepted.

Chapter 6

- **Crosse MJ**, ElShafei HA, Foxe JJ, Lalor EC (2015). Investigating the Temporal Dynamics of Auditory Cortical Activation to Silent Lipreading. *7th International IEEE/EMBS Conference on Neural Engineering*, 308–311.

Published Abstracts

- **Crosse MJ**, Lalor EC (2015). Eye Can Hear Clearly Now: Visual Speech Increases Sensitivity of Auditory Cortex to Speech in Peri-Threshold Noise. *Society for Neuroscience 2015*, Chicago, USA.
- **Crosse MJ**, Lalor EC (2015). Cortical Entrainment to the Speech Envelope during Audiovisual Speech Processing: a Correlated and Complementary Mode Perspective. *International Multisensory Research Forum 2015*, Pisa, Italy.
- **Crosse MJ**, ElShafei HA, Foxe JJ, Lalor EC (2015). Investigating the Temporal Dynamics of Auditory Cortical Activation to Silent Lipreading. *IEEE/EMBS Neural Engineering (NER) 2014*, Montpellier, France.
- **Crosse MJ**, ElShafei HA, Lalor EC (2014). Successful Lipreading of Silent Speech Strengthens the Correlation between Cortical Activity and the Corresponding Speech Envelope. *Auditory Cortex 2014*, Magdeburg, Germany.
- **Crosse MJ**, Lalor EC (2014). The Effects of Congruency on the Latency of Continuous Audiovisual Speech Processing. *International Multisensory Research Forum 2014*, Amsterdam, The Netherlands.
- **Crosse MJ**, Lalor EC (2013). The Effects of Audiovisual Speech Congruency on the Representation of the Speech Envelope in Human Auditory Cortex. *Society for Neuroscience 2013*, San Diego, USA.
- **Crosse MJ**, Lalor EC (2012). Multisensory Integration Effects to Continuous, Natural Audiovisual Speech with High Temporal Resolution. *Society for Neuroscience 2012*, New Orleans, USA.

Table of Contents

List of Figures

List of Tables

Glossary of Acronyms

A	– Audio-only
AEP	– Auditory Evoked Potential
ASD	– Autism Spectrum Disorder
AV	– Audiovisual
AVi	– Audiovisual incongruent
AVif	– Audiovisual incongruent female
AVin	– Audiovisual incongruent nature
AVsf	– Audiovisual static face
BCI	– Brain-Computer Interface
CDF	– Cumulative Distribution Function
CN	– Cochlear Nucleus
DCN	– Dorsal Cochlear Nucleus
ECoG	– Electrocorticography
EMG	– Electromyography
EEG	– Electroencephalography
ERF	– Event-Related Field
ERP	– Event-Related Potential
fMRI	– functional Magnetic Resonance Imaging
GFP	– Global Field Power
IC	– Inferior Colliculus
LTI	– Linear Time-Invariant
MEG	– Magnetoencephalography
MGN	– Medial Geniculate Nucleus
MSI	– Multisensory Integration
NRC	– Normalised Reverse Correlation

pSTS	– posterior Superior Temporal Sulcus
RMS	– Root Mean Square
RT	– Reaction Time
SC	– Superior Colliculus
SNR	– Signal-to-Noise Ratio
SI	– System Identification
SSAEP	– Steady State Auditory Evoked Potential
STG	– Superior Temporal Gyrus
STS	– Superior Temporal Sulcus
TRF	– Temporal Response Function
V	– Visual-only
Vk	– Visual known
Vu	– Visual unknown
VCN	– Ventral Cochlear Nucleus

Chapter 1 Introduction

> "The whole is greater than the sum of its parts."
> -Aristotle

When having a conversation in a noisy environment such as a busy restaurant, it is not always possible (or good manners) to lean in close to our interlocutor to better hear them. Instead, we instinctively look at their face to disambiguate the acoustic speech content. But how can looking at a person's face help us to better hear them if we are unable to lipread them in the absence of auditory speech? For thousands of years, this question has puzzled science philosophers, intrigued by how effectively the senses work together compared to in isolation. Indeed, this phenomenon is not limited to improving how we process speech stimuli. Our experience of the world is predominantly multisensory, thus encoding, decoding and interpreting biologically significant multisensory events are among the brains most important functions.

Of course there are obvious evolutionary advantages to being able to perceive our world through multiple senses. Our surroundings are rich with sensory information about important biological events such as prey, predators and mates. While individually, each sense is effective in specific circumstances, collectively, they increase the likelihood of detecting and identifying ecologically relevant events. However, the advantage of having multiple senses is further exploited when they are employed simultaneously, i.e., in a multisensory context. The integrated product of a multisensory percept reveals more about the nature of an event than would be predicted by the collective sum of its constituent components. As a result, we perceive events faster and more accurately, and can thus act more efficiently in our environment. This type of behaviour is ecologically significant for survival in any species. The improvement in behaviour observed as a result of multisensory processing arises from synergistic interactions at the neuronal level and is commonly referred to as multisensory integration.

Owing to recent developments in functional brain imaging techniques, it is now possible to non-invasively study multisensory integration in the human brain. In particular, audiovisual (AV) speech processing has become a popular theme in neuroscience research. Aside from furthering our understanding of how the human brain works, this growing field has practical applications in clinical research, as impaired multisensory processing has been implicated in numerous clinical populations. However, efforts to investigate the neural basis of AV speech integration have often focused on the special case of discrete AV tokens such as syllables or phonemes due to methodological constraints. Such stimuli are not fully reflective of natural speech, which is continuous, dynamical and rife with lexical constraint. Fortunately, recent electrophysiological studies have demonstrated that cortical response measures to continuous auditory speech can be easily obtained using system identification (SI) techniques.

Investigating the neural basis of natural AV speech processing (and to a lesser extent silent lipreading) using a newly-developed multisensory SI framework forms a common theme throughout the thesis. This chapter provides some background into the most relevant clinical research questions that motivate much of the research presented in the thesis and the context in which the succeeding chapters may be considered. In addition, the overall aims of the thesis are outlined and the chapter structure is described in brief.

1.1 Background

Multisensory integration is an autonomous neural process, allowing us to locate and identify objects more rapidly and accurately, and to perceive ecologically relevant stimuli such as speech more reliably. Aside from enhancing our perception of the world, it also allows us to make sense of the plethora of multisensory events bombarding our senses in everyday life. While most of us take this for granted, we would likely find it difficult to function in society if we could not integrate all of this information into coherent, meaningful percepts. For example, if we were unable to integrate multiple sensory inputs, then our environment, which by nature is multisensory, would become a complex and confusing space. As a result, we would be unable to make sense of this space, becoming overwhelmed and withdrawing from it. This idea has led researchers to believe that impaired multisensory processing may be a core deficit in neurodevelopmental disorders such as autism spectrum disorder (ASD; Foxe and Molholm, 2009, Donohue et al., 2012). While multisensory deficits have also been reported in other clinical disorders such as dyslexia (Hairston et al., 2005) and schizophrenia (Ross et al., 2007b, Stekelenburg et al., 2013), its utility as a ‘biomarker’ of ASD has received considerable attention. This is because ASD onsets at a very early stage in childhood and developmental outcome can decline significantly if intervention does not occur early enough.

ASD is characterised by deficits in social interaction, communication and the presence of restricted and repetitive behavioural patterns (APA, 2013). It is estimated that the disorder affects approximately 1 in 68 children aged 8 years old in the United States (Wingate et al., 2014) and is frequently comorbid with other psychiatric conditions, most notably attention-deficit/hyperactivity disorder (Skokauskas and Gallagher, 2012). Autism comes at a high cost to the individual, such that it results in poorer employment prospects and reduced societal engagement. Developmental outcome can, however, be improved if intervention is provided prior to the age of two years (Dawson et al., 2010). Intervention at a young age is hindered by the difficulty often encountered in trying to secure a reliable diagnosis of the condition where the child’s verbal communication skills are not fully developed, as many of the current clinical assessments rely on subjective report of symptoms in addition to third-party reports. Thus, it has long been acknowledged that an objective physiological measure of symptom and deficit severity is necessary to facilitate early diagnosis and intervention in ASD.

While considerable phenotypic variation exists in ASD, a common deficit across the spectrum is atypical behavioural responses to sensory stimuli, with over 96% of autistic children reporting abnormal sensitivity across multiple sensory modalities (Dunn et al., 2002). Particular sensory stimuli have been shown to induce self-injurious and aggressive behaviour in children with autism (Leekam et al., 2007). This emerging body of evidence strongly suggests that sensory impairments may be at the very core of the disorder, such that they were recently included in the DSM-5 criteria for diagnosing ASD in addition to the well-established language, communication, and social deficits (APA, 2013). This has resulted in a deluge of research examining the neurophysiological correlates of sensory deficits in autism (for reviews, see Marco et al., 2011, McDevitt et al., 2015). It is considered that understanding the neurobiological processes associated with impaired sensory processing in ASD will give rise to neural biomarkers of the disorder that could be used to facilitate earlier diagnosis and intervention, and to monitor the impact of different therapeutic strategies.

There is strong empirical evidence to suggest that multisensory integration is impaired in children with ASD (Russo et al., 2010, Brandwein et al., 2013, Stevenson et al., 2014b). This has been demonstrated at both a low and high level of AV processing (Marco et al., 2011) and it is significantly more pronounced for high-level (linguistic) AV processing (Bebko et al., 2006). A recent study tracing the developmental trajectory of this impairment in AV speech processing demonstrated that it is more pronounced in younger children (7–9 years) than in adolescents (13–16 years; Foxe et al., 2015). In a recent electrophysiological study (Brandwein et al., 2015), it was demonstrated that multisensory integration of low-level AV stimuli was altered in children with ASD and that these electrophysiological indices were associated with symptom severity. Electrophysiological differences have also been shown in adolescent males using higher-level stimuli such as AV speech (Megnin et al., 2012).

Despite the potential utility of AV speech to probe electrophysiological markers of ASD in children, it has not received the proportionate research interest. It is also possible that AV speech integration in ASD has not been probed under the right environmental conditions – the benefit of multisensory speech is much greater when the speech is presented in noisy and distracting environments (Ross et al., 2007a) and it is precisely these environments that present the greatest difficulties for individuals with ASD. Furthermore, the aforementioned limitations in traditional brain imaging analysis methods have meant that the majority of electrophysiological research on multisensory speech has examined the brain’s response to discrete, unnaturalistic AV stimuli. The development of an SI approach for studying multisensory integration using naturalistic AV speech stimuli would therefore have important implications for furthering the utility of electrophysiological research in ASD.

1.2 Aims

The overall aim of the thesis is to develop an analysis framework for studying how the human brain processes and integrates natural, continuous AV speech using electrophysiological recordings. The project has three core aims:

1. To develop a MATLAB-based SI toolbox for studying sensory processing that is easy to use and accessible to the broader neuroscience community.
2. To establish an SI framework specifically for quantifying multisensory integration in natural AV speech processing.
3. To further our understanding of the neural mechanisms that underpin the manner in which the human brain processes and integrates natural AV speech.

The overarching objective of the thesis is clear and builds incrementally through the implementation of the specific aims outlined above.

1.3 Thesis Outline

In Chapter 2, the anatomy of speech processing is introduced, with particular focus on the auditory system. A brain imaging technique called electroencephalography (EEG) and its application in studying speech processing is discussed, as it is used in all the studies described in the thesis. Important background information on the role of AV speech is also provided, leading into a detailed description of multisensory integration in the context of AV speech processing. Particular focus is given to behavioural and neurophysiological studies that have investigated AV speech processing using discrete, syllabic stimuli. The final section discusses neurophysiological studies that have investigated speech processing in a more naturalistic way, as well as several AV speech studies that have exploited such an approach.

Chapter 3 begins by introducing the technique of SI in the context of sensory neuroscience. The remainder of the chapter is devoted to a description of a MATALB-based toolbox, called mTRF Toolbox, which was developed as part of this research work to enable easy implementation of an SI approach for studying sensory processing. The mathematical background underlying the approach is described in full, followed by practical advice on the implementation of the toolbox. Concrete EEG examples are given that demonstrate the usage and versatility of the toolbox, as well as a comparison with traditional methods. Finally, applications and important considerations of the toolbox are discussed in the context of sensory neuroscience.

In Chapter 4, a framework is developed for quantifying multisensory integration in natural AV speech processing using the mTRF Toolbox. Specifically, this study uses EEG to examine the role of contextual and temporal congruency in AV speech integration. Several carefully designed AV speech conditions are implemented to probe potential neural mechanisms underpinning AV speech integration in quiet listening conditions. The mTRF Toolbox is employed in several complementary ways to address these questions.

Chapter 5 builds on Chapter 4 by examining AV speech integration in degraded listening conditions using spectrally-matched noise. The paradigm is designed such that the behavioural benefit of multisensory processing is maximised relative to unisensory processing. The data from Chapter 4 are also reanalysed and compared with these new data to provide a reference with which inverse effectiveness can be examined. Using the mTRF Toolbox, new analysis techniques are explored to elucidate the neural mechanisms specific to AV speech integration in adverse hearing conditions. The relationship between our neural and behavioural indices of multisensory integration is also examined.

Chapter 6 investigates the nature of auditory cortical activation to silent lipreading. Specifically, different levels of lipreading performance are examined to probe the impact that it has on the temporal dynamics of cortical activation in auditory regions. The utility of the mTRF Toolbox for decoding acoustic information during silent lipreading is discussed in the context of brain computer-interface (BCI) applications.

Research in the field of multisensory integration has grown over the past number of decades, particularly since the inception of the annual International Multisensory Research Forum (IMRF) conference in Oxford in 1999 (Foxe and Molholm, 2009). Much of this research has begun to focus on AV speech processing because of its clinical relevance and application, not to mention its ecological importance to humans. While many of these studies have yielded valuable insight into how the human brain integrates multisensory information, it will surely require the development of new methodological approaches to further the impact of this work on the field. It is hoped that the studies presented in this thesis will make a significant scientific contribution to the field of multisensory integration and, in their application, to certain clinical fields, including ASD.

Chapter 2 The Electrophysiology of Audiovisual Speech Processing

This chapter provides a review of the literature relevant to this thesis and is divided into five sections. The first section provides an introduction to the human brain, in particular, the human auditory system. The second section describes the brain imaging technique, EEG, which is employed in all studies in the thesis. The third section discusses the role of visual cues in speech comprehension. The fourth section presents a summary of previous behavioural and neurophysiological research on multisensory integration in the context of AV speech. The final section gives an EEG account of AV speech integration, concluding with recent advances in methodological techniques.

2.1 The Anatomy of Speech Processing

2.1.1 The Cerebral Cortex

The adult human brain accounts for almost 97 percent of the body's neural tissue (Martini and Nath, 2009). The brain consists of several principal structures, each with specific functions. The cerebrum, which dominates most of the brain's mass, can be divided into two cerebral hemispheres. Each of these hemispheres is subdivided into four lobes: frontal, parietal, occipital and temporal (Bear et al., 2007). The outer layer of the human cerebrum is called the cerebral cortex and is highly folded and covered in a superficial layer of grey matter known as neocortex. The cerebral cortex forms a series of elevated ridges known as *gyri* which are separated by shallow depressions known as *sulci* or by deeper grooves known as fissures. These fissures bound the different aforementioned regions of the cerebrum. The two cerebral hemispheres are almost completely separated by a deep interhemispheric or longitudinal fissure. On each of the hemispheres, a deep groove known as the central sulcus divides the anterior frontal lobe from the more posterior parietal lobe. The frontal lobe is separated from the more inferior temporal lobe by the lateral sulcus (known as the Sylvian fissure) and from the more posterior parietal lobe by the central sulcus (Bear et al., 2007). The parietal lobe is separated from the even more posterior occipital lobe by the parieto-occipital sulcus.

Each region of the cerebral cortex is defined not only by its location, but also by its functionality, although the entirety of this functionality is still not well established. The primary sensory areas, which are the first to receive signals from ascending sensory pathways, are each located within a different cerebral lobe (Martini and Nath, 2009). The primary somatosensory cortex is located in the parietal lobe and is responsible for our conscious perception of touch, pressure, pain, vibration, taste and temperature. Primary auditory cortex (A1) is located in the temporal lobe, while primary visual cortex (V1) is located in the occipital lobe. Information from these primary sensory areas is then passed on via millions of interconnections to secondary sensory areas where it is further processed. A third region of cortex consists of motor areas, which are concerned with voluntary contraction of skeletal muscles. The primary motor cortex is situated along the precentral gyrus of the frontal lobe, just anterior of the central sulcus. Distinct areas of the motor cortex control specific parts of the body separate from those governing other parts. This spatial arrangement of functionality also occurs in the auditory cortex (tonotopic organisation) and the visual cortex (retinotopic organisation). The sensory and motor areas are further connected to large regions of cortex known as association areas. These areas are responsible for higher-level cognitive processing such as interpretation of sensory input and motor response coordination.

Regions of auditory, visual and even motor cortex are all involved in the processing of AV speech stimuli. However, as the main theme of the thesis is how visual speech impacts on auditory speech processing (i.e., how the acoustic signal is processed), the following section is devoted to a description of the auditory system. The reader is directed to Martini and Nath (2009) for a more complete description of the anatomical organisation of the visual and motor cortices.

2.1.2 The Auditory System

The Peripheral Auditory System

The peripheral auditory system can be split into three components: the outer ear, the middle ear and the inner ear (see Fig. 2.1; Purves et al., 2008). Hearing begins at the outer ear where incoming sound waves are funnelled into the ear canal by folds of cartilage and skin known as the *pinna* or *auricle*. This intricately designed organ also allows us to locate sounds in three-dimensional space. The ear canal is the resonant cavity located between the outer and middle ear. It usually has a resonant frequency of around 2–5 kHz, thus amplifying sounds within this frequency range. Much of the frequency content of speech is contained within this range which is partly why humans have a heightened sensitivity to speech (Robinson and Hawksford, 1999). The ear canal also increases the sound pressure level by up to 20 dB at certain frequencies. Once the sound wave has been funnelled down the ear canal, it reaches the *tympanic membrane,* i.e., the ear drum. The pressure of the sound wave impinges on the membrane causing it to vibrate.

Beyond the ear drum is an air filled cavity known as the middle ear. This contains the auditory ossicles which are the three small bones known as the *malleus*, *incus* and *stapes* (Fig. 2.1). They are the interface between the outer and inner ear, forming a conductive chain from the tympanic membrane to the oval window of the *cochlea*. The ossicles are a mechanical system that acts as an impedance matching network, allowing for efficient energy transfer between the outer and inner ear. Specifically, they transform a relatively large displacement and small force at the ear drum to a small displacement and large force at the oval window.

The inner ear represents the interface between the auditory and nervous system and contains a small coiled tube known as the cochlea (Fig. 2.1). The cochlea essentially performs frequency analysis on the incoming mechanical signal by splitting it up into multiple frequency bands (Yang et al., 1992). A complex structure within the cochlea known as the *basilar membrane* determines the mechanical wave properties of the cochlea. The cochlea is filled with an incompressible liquid and movement of the ossicles against the oval window causes a hydrostatic pressure which in turn sets up a travelling wave in the cochlear fluid, propagating from the base towards the apex of the basilar membrane, growing in amplitude and slowing in velocity until a point of maximum displacement is reached (Purves et al., 2008). High frequencies maximally displace the base of the membrane, whilst low frequencies maximally displace the apex, giving rise to a tonotopic organisation (i.e., frequency-to-place mapping). A structure known as the *organ of Corti* contains tiny hair like structures which move in response to any deformation in the basilar membrane. These 30,000 or so hair cells are connected to nerve cells which generate neural signals when they detect movement (Rice, 2009).

Essentially, the cochlea acts like an analogue filterbank, splitting the soundwave into logarithmically-spaced frequency bands and outputting the rectified signal intensity at each band. The intensity of the signal at each frequency band is determined by how many of the hair cells in the cochlea are stimulated. Slow (<50 Hz) modulations in intensity in each frequency band are known as the narrowband envelopes and their summation across all frequency bands is known as the broadband envelope (Rosen, 1992). The envelope of speech can convey important segmental cues to a variety of linguistic information such as manner of articulation, voicing, vowel identity and prosodic cues (Rosen, 1992). Most of the information in the speech envelope is at frequencies below ~8 Hz (Houtgast and Steeneken, 1985). However, envelope frequencies critical for speech comprehension are contained between 4–16 Hz (Drullman et al., 1994, van der Horst et al., 1999), and some above 16 Hz (Shannon et al., 1995). Thus, frequency analysis that is critical for extraction of the speech envelope (which is critical for linguistic processing) has already begun by the time the speech signal has left the inner ear.

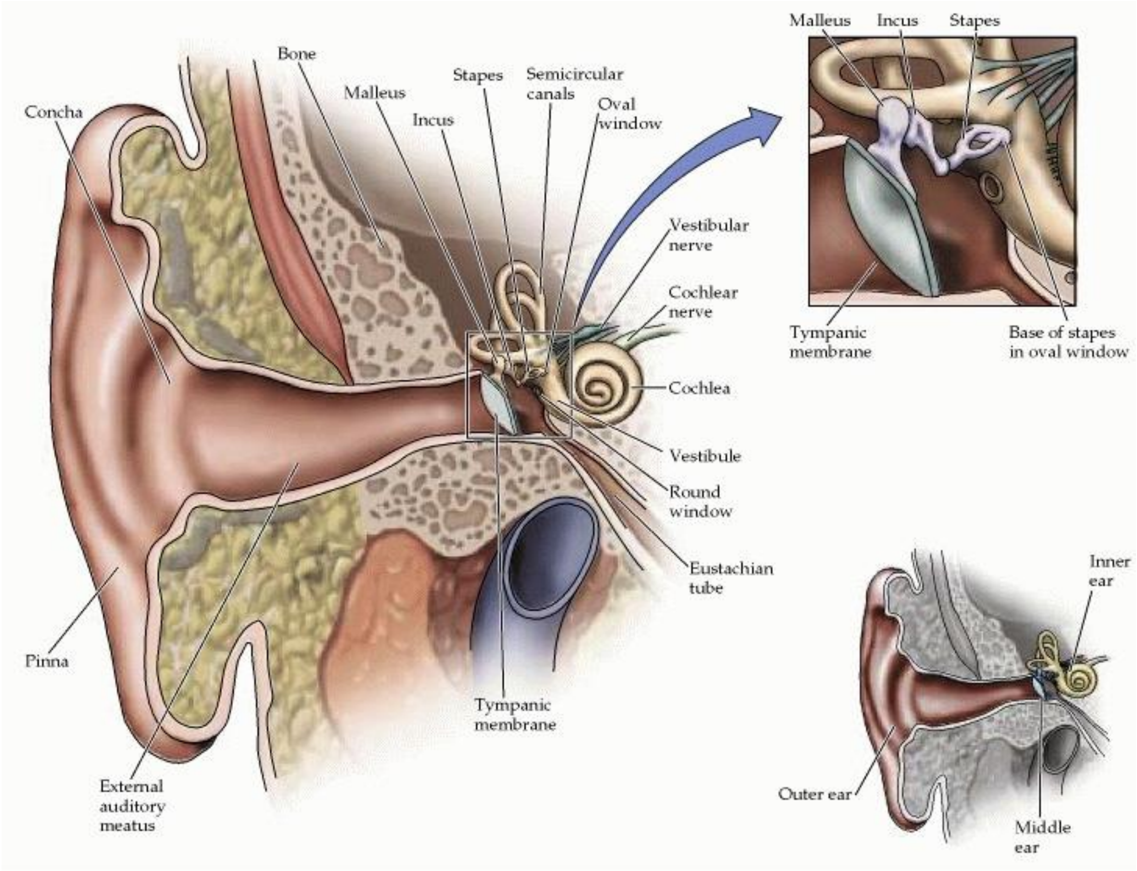


Figure 2.1: The auditory periphery (Purves et al., 2008).

The Central Auditory System

The inner ear is connected to the central auditory system via the *vestibulocochlear nerve* which, as its name suggests, consist of a vestibular branch and a cochlear branch. As the vestibular branch is not concerned with auditory processing, it will not be considered further in the thesis. The cochlear branch (i.e., the auditory nerve) monitors the receptors in the cochlea and carries information concerned with hearing. These axons enter the brainstem at the *cochlear nucleus* (CN), a structure that preserves the tonotopic organisation established by the cochlea. The CN can be subdivided into the left and right *dorsal* and *ventral cochlear nuclei* (DCN and VCN respectively). The VCN extracts information on the firing rate and population activity of the auditory nerve fibres, while the DCN performs nonlinear spectral and spatial analyses (Purves et al., 2008).

Immediately after the CN, auditory information is projected laterally to the *superior olivary complex*. Here, for the first time, information from both ears converges, allowing analysis related to binaural hearing to be computed. Specifically, interaural time difference and interaural level difference are computed which determines the direction from which a sound originates in the azimuth (left/right) plane. From there, information ascends to the *inferior colliculus* (IC) of the mid brain (see Fig. 2.2; Purves et al., 2008), a centre that directs a variety of unconscious motor responses to sounds. It is thought that neuronal coding of auditory information is transformed in IC and representational maps of neuronal response features are formed, from which perception of sound may be derived (Ehret and Romand, 1997). These ascending auditory signals then synapse in the *medial geniculate nucleus* (MGN) of the thalamus, which acts as a relay between IC and auditory cortex. Please refer to Purves et al. (2008) for further reading on the peripheral and central auditory systems.

Auditory Cortex

The auditory cortex is located in the superior portion of the temporal lobe, mostly hidden within the lateral sulcus. The auditory cortex can be subdivided into three regions: the *core, belt* and *parabelt*. The core is located deep within the lateral sulcus and receives input from MGN via the *superior temporal gyrus* (STG) or Heschl's gyrus. Like the cochlea, the core is tonotopically organised, although it is thought that the majority of low-level (spectrotemporal) acoustic processing has already been carried out by the time the signal reaches it (Nelken, 2008). The belt is a narrow band of cortex that surrounds the core and also receives input from MGN as well as the core. The belt is less responsive to pure tones but exhibits some tonotopic organisation. Belt regions are highly interconnected and project primarily to the parabelt. The parabelt adjoins the lateral belt area and receives input from the MGN as well as the belt.

From there, the parabelt projects to several regions in the frontal lobe, as well as portions of the parietal and temporal lobes. One such region of particular interest, the *superior temporal sulcus* (STS), will be discussed in detail in section 2.4.4. Speech, which is processed primarily in auditory cortical regions, can be organised into a hierarchy of perceptual units, i.e., phonemes, words, sentences, etc. (Chomsky and Halle, 1968). Linguistic processing is thought to start in posterior STG with phonemic analysis, followed by the formation of words in middle STG, eventually projecting to anterior STS, where a sentential representation is formed (see meta-analyses in DeWitt and Rauschecker, 2012, Davis and Gaskell, 2009, Adank, 2012). However, it has since been demonstrated that STG topographically encodes phonetic features, and not individual phonemes (Mesgarani et al., 2014). While it is widely considered that language is predominantly processed by most people in the left hemisphere (Hickok and

Poeppel, 2007), current perspectives suggests that it is processed bilaterally in the early stages of linguistic processing, but becomes more and more left-lateralised further up the auditory hierarchy (Peelle, 2012).

Researchers employ a wide variety of functional brain imaging techniques to interrogate how the brain processes speech. As all of the studies in this thesis employed the method of EEG, the next section is devoted to a description of its functional basis. For a more complete discussion on how language is processed in the brain, please refer to (Hickok and Poeppel, 2007).

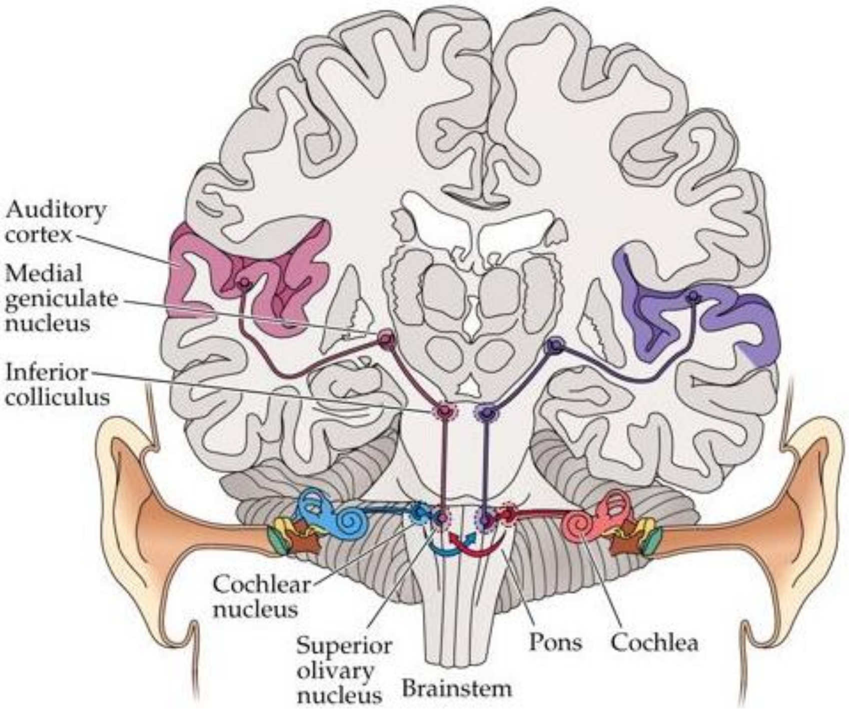


Figure 2.2: The ascending auditory pathway (Purves et al., 2008).

2.2 Electroencephalography

The brain contains billions of neurons and their activity elicits electric potentials that can be measured from the scalp surface using EEG. These potentials are primarily generated by a particular type of cortical cell known as a pyramidal cell. Pyramidal cells account for approximately 80 percent of all cortical cells and most of these are orientated perpendicular to the cortical surface (Bok, 1959). These cells consist of a set of branch-like dendrites that receive input from other neurons, a cell body, and an axon, which delivers electrochemical output to receiving neurons (see Fig. 2.3*A*). The point of connection between two neurons is called the synapse. The electrochemical output is an electrical response of fixed duration and amplitude known as an *action potential*. This response works on an all-or-nothing basis, i.e., an action potential will only propagate along a cell's axon if the action potential received by the neuron caused sufficient change in membrane potential at the cell body. This synaptic activity causes current flow which in turn produces an electric field (shown in Fig. 2.3*A*). This field pattern resembles that produced by a dipole source at distances larger than a few lengths of the neuron (Nunez and Srinivasan, 2006). The synchronous activity of these strongly interconnected cells is essentially what produces EEG activity at the scalp.

The potentials generated by these cells can be used to build an image of the neural activity across the brain's surface in response to an event such as speech. Due to the use of surface electrodes, spatial resolution is quiet limited; a high-density EEG system can have up to 512 electrodes which relates to an inter-electrode distance of approximately 11 mm (Gevins et al., 1991). Even this level of electrode density is uncommon in research. For instance, the studies presented here employed 128-channel EEG, which has four times less spatial resolution. Each electrode is strategically positioned according to the standardised 'International 10–20 System' (Fig. 2.3*B*; Jasper, 1958) and measures potentials generated by approximately 10^7 to 10^9 neurons (Nunez, 1995). The electric fields elicited by cortical neurons must first pass through several anatomical layers, including cerebrospinal fluid, the meningeal layers (the *dura*, *arachnoid*, and *pia*), skull bone, periosteum and skin tissues, before reaching the electrode surface. These materials act as spatial filters and attenuate the signal being recorded. A high level of amplification is therefore required to detect the signal, which is in the order of 20–40 μV. Unfortunately, this also amplifies other unwanted signals of greater magnitude such as electrooculogram activity from eye blinking or movement (Corby and Kopell, 1972) and electromyogram (EMG) activity from muscle activation (Goncharova et al., 2003). The layers of tissue between the brain and the electrodes also cause the EEG response to smear across the scalp, an effect known as 'volume conduction' (Freeman et al., 2003). This further reduces the precision with which neural sources can be localised within the brain.

What EEG lacks in terms of signal-to-noise ratio (SNR) and spatial resolution, it makes up for in temporal resolution, which is on the order of milliseconds rather than seconds. EEG is typically recorded at sampling rates between 250 and 2000 Hz, but is capable of recoding at sampling rates above 20 kHz if necessary. However, the aforementioned anatomical layers (i.e., cerebrospinal fluid, meninges, skull, scalp) also act as low-pass filters, attenuating important neural information, particularly that above ~30 Hz. Fortunately, activity from cortical neurons has been shown to track the temporal envelope of natural speech below ~10 Hz (Luo and Poeppel, 2007, Ding and Simon, 2012b). Thus, both the temporal resolution and bandwidth of EEG make it a very suitable method for investigating how the brain processes speech and, in particular, how it tracks the envelope of speech.

Other brain imaging techniques that will be referred to in this thesis include: electrocorticography (ECoG), magnetoencephalography (MEG) and functional magnetic resonance imaging (fMRI). ECoG is the most similar technique to EEG in that it measures electric potentials elicited by the brain. However, these measures are taken directly from the brains surface or in even in single neurons making it a highly invasive procedure. Hence, this can only be performed in per-surgical epileptic patients, making it inaccessible to most researchers. MEG is also similar to EEG in that it measures the magnetic counterpart of the electric fields elicited by neurons and is non-invasive. While it affords spatial resolution superior to EEG, it is far more expensive and not readily portable. fMRI, on the other hand, is a neuroimaging technique that measures changes in blood flow within the brain, thought to reflect the energy used by neighbouring cells, i.e., neuronal activity. This technique offers the best spatial resolution but very poor temporal resolution (~1 Hz). Furthermore, it is highly expensive, immobile and extremely noisy. For a more complete overview of functional brain imaging techniques, please refer to Papanicolaou (1998).

The electrical activity measured by EEG can be organised into two main categories: (1) event-related potentials (ERPs), elicited in response to any sort of discrete time-locked event or sensation, and (2) spontaneous oscillatory or rhythmic potentials, which occur naturally in the brain during both stimulation and resting state. The next two sections provide descriptions of these types of EEG activity and their role in speech research.

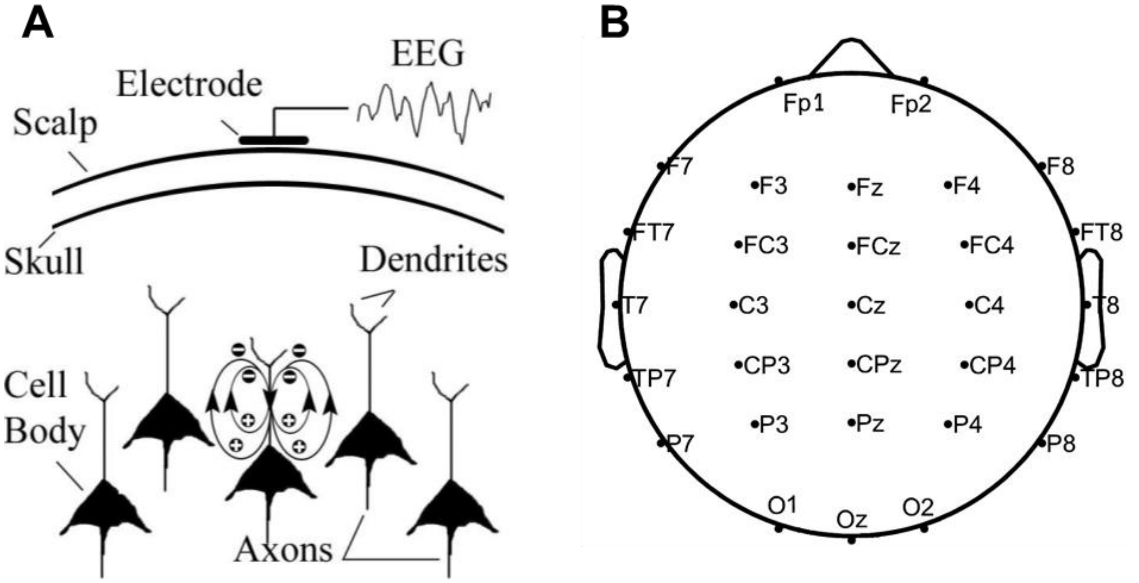


Figure 2.3: Measuring neuronal population activity with EEG.
A, Orientation of pyramidal cells in the outer cortex. ***B***, The international 10–20 system for electrode placement (adapted from Sakkalis et al., 2010).

2.2.1 Event-Related Potentials and the Auditory Evoked Potential

One of the most common methods for utilizing EEG to study how our brains process sensory stimuli is to examine ERPs. When a stimulus is presented, different neuronal populations are activated in series and in parallel. This sequence of polarising and depolarising cell membranes generates a fluctuating potential which give the ERP its characteristic trace. These deflections represent the sum of several relatively independent components which are difficult to isolate and measure independently (Handy, 2005). However, in recording situations it is even difficult to identify entire ERPs because they are generally much smaller than spontaneous EEG, i.e., noise. By assuming noise has a zero mean, an averaging technique can be used to eliminate it and preserve just the response to the stimulus. This is achieved by extracting a series of EEG epochs, which are time-locked to a repeated stimulus, and averaging them. The MEG equivalent is known as an event related field (ERF) and is calculated in much the same way. In terms of AV speech, most EEG-based research has focused on the ERP elicited in response to the auditory modality, known as the auditory evoked potential (AEP; Davis, 1939). This response can be subdivided into three sequences of waves representing activity among different cell populations of the auditory hierarchy. The first two sequences are known as the brain stem response and middle-latency sequence, and occur between 8–12 ms and 40–50 ms post-stimulus, respectively (see Fig. 2.4*A*;

Picton, 2013). The latter is thought to be generated by activity in the thalamus and primary auditory cortex (Picton et al., 1974).

The final sequence is of most interest here as it relates to speech-specific processing. Known as the long-latency or cortical response, it occurs between 50–500 ms and is thought to reflect activity in higher-order auditory and association cortex. This sequence is made up of two positive peaks (P1 and P2), typically occurring at around 50 and 150 ms respectively, and two negative peaks (N1 and N2), typically occurring at around 100 and 200 ms respectively (Fig. 2.4*A*; Picton, 2013). These response components have been linked to the different stages of linguistic processing outlined in section 2.1.2 (Salmelin, 2007, Picton, 2013). Specifically, it is thought that acoustic-phonetic analysis of speech starts around 50–100 ms, generating the P1 response in primary auditory cortex and the N1 response in non-primary auditory cortex. Language-specific phonetic-phonological analysis is thought to occur between 100–200 ms, generating the P2 and N2 response components. From 200 ms onwards, lexical-semantic processing is thought to occur in superior temporal regions (Fig. 2.4*B*; Salmelin, 2007).

Thus, the temporal profile of AEPs can be useful in studying how the brain processes speech. The next section outlines the potential role of cortical oscillations in speech processing based on their hierarchically-structured temporal scale.

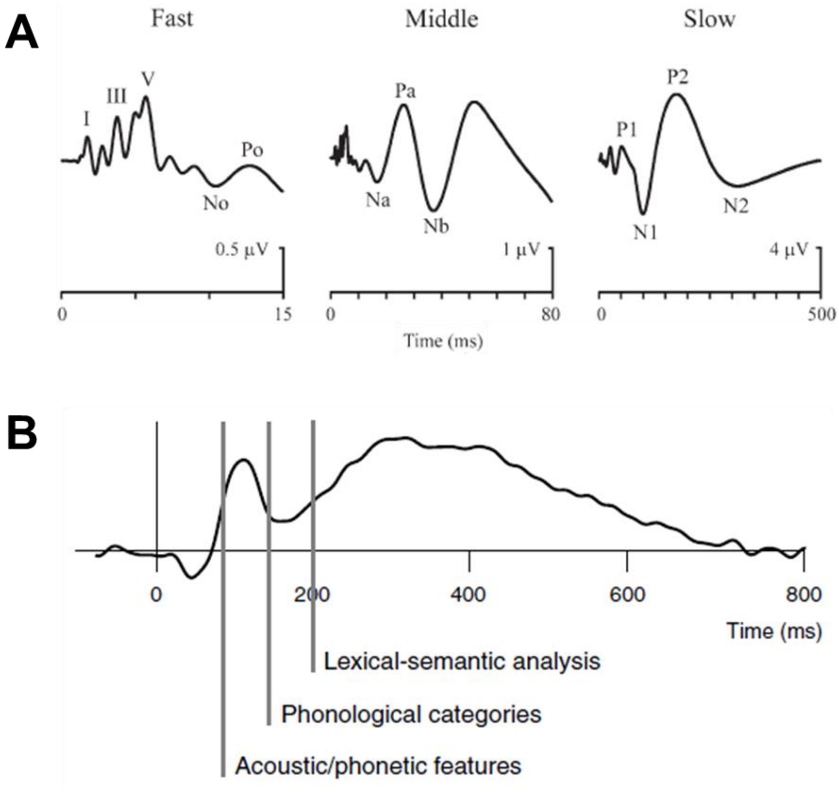


Figure 2.4: The temporal profile of the auditory evoked potential (AEP).

A, Early-, middle- and long-latency ERP responses (adapted from Picton, 2013). ***B***, Timecourse of speech processing in the superior temporal cortex (adapted from Salmelin, 2007).

2.2.2 Neural Oscillations and Speech Processing

In addition to transient stimulus-evoked responses, the cerebral cortex generates oscillatory activity that can be elicited spontaneously (ongoing oscillations) or by an external stimulus (evoked and induced oscillations). Spontaneous oscillations have been associated with particular states of behaviour and are categorised by the frequency at which they occur (see Fig. 2.5; Kent, 2010). Gamma (γ) rhythms are the fastest, oscillating at >30 Hz and are thought to represent higher-order processes such as consciousness, perception and problem solving. Beta (β) oscillations occur at 12–30 Hz and are thought to signal an active cortex. Alpha (α) rhythms are associated with relaxed, waking states and oscillate at 8–12 Hz. Next are theta (θ) oscillations, which occur during some sleep states at 4–7 Hz. Finally, delta (δ) rhythms are the slowest, oscillating at <4 Hz, but have the largest amplitude. These signals are thought to indicate deep sleep (Bear et al., 2007, Steriade et al., 1990).

Evoked oscillations, on the other hand, entrain to the phase of periodically modulating stimuli. If an auditory stimulus modulates at a constant frequency, the oscillation elicited is known as a steady-state auditory evoked potential (SSAEP) and will exhibit enhanced power at that specific frequency. If the stimulus is not perfectly periodic, like for example natural speech, then the evoked oscillation will entrain more loosely to the stimulus spectrum (Luo and Poeppel, 2007). A recent study demonstrated that if speech is presented in an artificially periodic manner, it elicits an SSAEP at the rate that the words were presented (Ding et al., 2016). More interestingly however, they showed that if the words can be grouped into phrases and then sentences with periods that are multiples of the words, then they two can induce lower-frequency SSAEPs that cannot be induced in a non-native speaker.

While evoked oscillations certainly play a major role in speech processing, recent perspectives have also suggested an active role for the naturally-occurring spontaneous oscillations mentioned earlier (Giraud and Poeppel, 2012). Previous work has shown that oscillations in auditory cortex are hierarchically organised in a way that structures its temporal activity so as to optimise the processing of rhythmic sensory inputs such as speech (Lakatos et al., 2005). Specifically, they demonstrated that delta phase modulates theta amplitude and that theta phase modulates gamma amplitude. The perceptual units of connected speech can be organised in a similar hierarchical structure (Poeppel, 2003). Thus, it is thought that low delta oscillations (1–2 Hz) could serve to parse the speech signal in a suprasegmental manner, i.e., into lexical and phrasal units. Theta oscillations (4–7 Hz) could parse smaller segmental units such as syllables, relying on information from the temporal envelope. Finally, low-gamma oscillations (30–50 Hz) could parse more fine-grained information at the phonetic scale (such as formant transitions), relying on the signal fine structure (Giraud and Poeppel, 2012). This parsing of segmental units is thought to be underpinned by a specific neural mechanism related to oscillations, known as phase-resetting (Hari and Salmelin, 1997, Engel et al., 2001). This theory posits that thalamocortical input resets the phase of theta oscillations such that the arrival of upcoming auditory information coincides with a high excitability phase of the auditory neuronal population, thus optimising segmentation of syllabic units (Luo and Poeppel, 2007). Because theta and gamma generators are coupled or nested together (Lakatos et al., 2005), this in turn resets the phase of gamma oscillations, initiating phonemic analysis (Giraud and Poeppel, 2012). While it has been proposed that ERPs could actually be the result of phase-resetting as opposed to the summation of evoked responses (Sauseng et al., 2007), both mechanisms will be considered within the context of the thesis.

Indeed, temporal processing is undoubtedly a critical part of how our brains analyse connected speech. EEG, with its high temporal resolution, is certainly well suited to studying these processes. As the focus of this thesis is how visual speech impacts the processing of acoustic speech, the next section is devoted to a description of multisensory speech and the role played by visual speech.

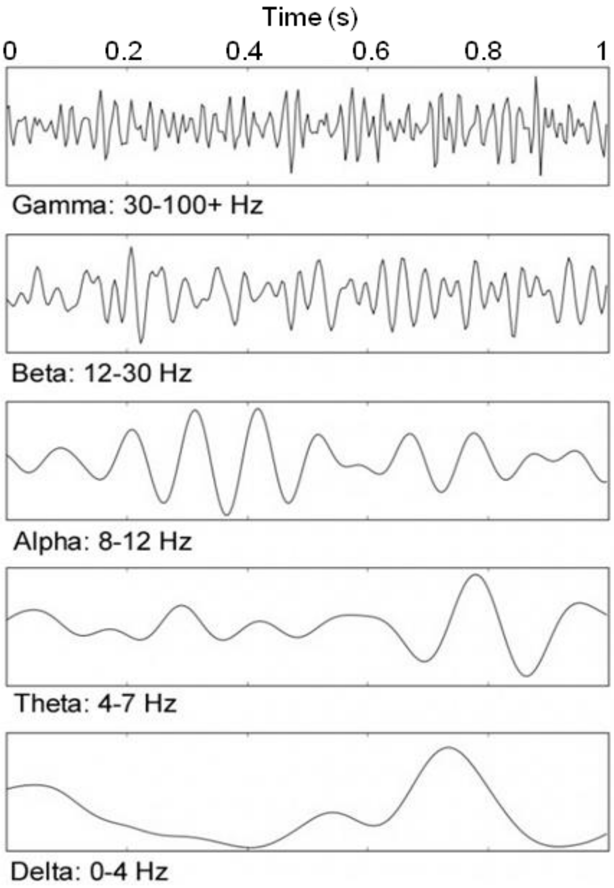


Figure 2.5: Cortical oscillations in EEG (adapted from Kent, 2010).

2.3 Human Speech: A Multisensory Experience

It is often overlooked in the study of human sensory systems that our experience of the world is profoundly multisensory and it is likely that multiple temporally overlapping sensory systems enable us to process these multimodal events seamlessly (Smith and Gasser, 2005). One complex class of multisensory signals that has received increasing attention is audiovisual speech, i.e., where a listener can both hear and see their interlocutor. In most social contexts, AV speech is the primary mode of speech perception and should not be thought of as a sub-class of auditory speech perception (Rosenblum, 2005). By definition, human speech is bimodal, dynamic and forms a large proportion of the sensory signals encountered by humans in everyday life (Chandrasekaran et al., 2009).

2.3.1 The Relationship between Auditory and Visual Speech

When viewing a speaker face-to-face, a wide range of useful information is available to the listener in the form of articulator, facial, head and hand movements. Identifying which of these signals (or which combinations of them) are most important to speech reading has been a topic of recent study (Yehia et al., 2002, Jiang et al., 2002). Furthermore, these signals are embedded within a rich statistical structure that is highly correlated with the spectrotemporal dynamics of the acoustic signal. It has been suggested that the human brain has evolved to encode these statistical regularities in the most efficient way possible by exploiting redundancies and correlations of the input space (Barlow, 1961).

Several studies have examined the spectrotemporal characteristics of AV speech, demonstrating that the area of the mouth opening and the broadband envelope of the acoustic speech signal are highly correlated with each other (Fig. 2.6; Chandrasekaran et al., 2009, Grant and Seitz, 2000). This correspondence was shown to be most robust for the envelope extracted at frequencies below 1 kHz and between 2–3 kHz, commensurate with formant frequencies F1 and F2–F3 respectively. Chandrasekaran et al. (2009) also demonstrated that both the broadband envelope and mouth area are temporally modulated in the 2–7 Hz range, which overlaps with the timescale of the syllable (Kuwabara, 1996). Furthermore, they found that the onset of mouth movements consistently preceded the onset of vocalisations by between 100 and 300 ms. This particular finding led a number of neuroscience papers to assume that visual speech consistently leads auditory speech by ~150 ms (Arnal et al., 2009, Arnal et al., 2011, Zion-Golumbic et al., 2013a). However, a recent study by Schwartz and Savariaux (2014) highlighted that the temporal relationship between auditory and visual speech is indeed more complex, consisting of a range of AV asynchronies that vary from small audio leads (20–40 ms) to large audio lags (70–200 ms).

Moving beyond the kinematics of the mouth, research has shown that facial movements, tongue movements and speech acoustics are predictive of each other at the level of syllables as well as sentences (Jiang et al., 2002). A study which attempted to predict acoustic speech patterns from the available visual information found that mouth and facial movements (particularly eyebrows) contributed more to the synthesis than the movements of just the mouth and lips (Yehia et al., 2002). In line with this, it has been shown that we are more inclined to focus on a speakers eyes than on their mouth in low levels of background noise, and that we only become more biased towards the mouth in high levels of noise (Vatikiotis-Bateson et al., 1998). It has also been demonstrated that the kinematics of the head are temporally aligned to the spectrotemporal dynamics of the speaker's voice, conveying suprasegmental features of speech such as stress and prominence, i.e., prosody (Munhall et al., 2004a).

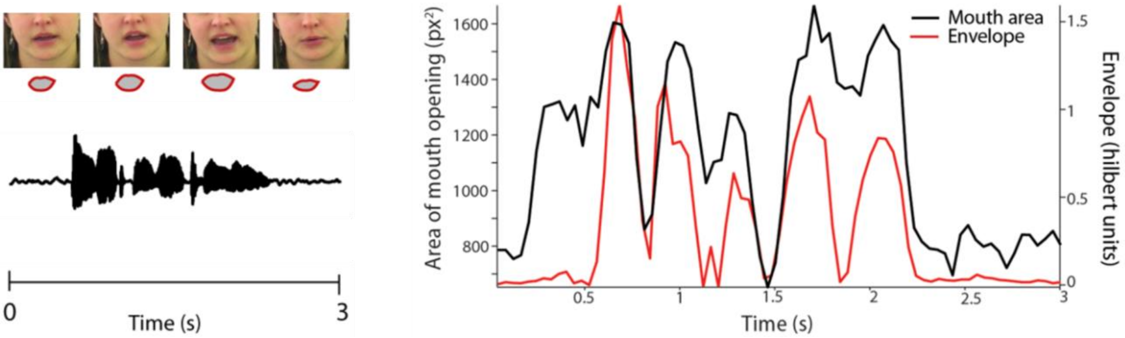


Figure 2.6: Temporal relationship between mouth area and the acoustic envelope (adapted from Chandrasekaran et al., 2009).

2.3.2 Defining the Role of Visual Speech

Current perspectives on AV speech argue that it can be characterised in terms of two specific modes of multisensory information: 'correlated' and 'complementary' (Campbell, 2008, Summerfield, 1987, Grant and Seitz, 2000). These modes of speech are defined by the role that visual cues play in improving speech comprehension in noise, in those with impaired hearing and in the case of ambiguous speech features.

Visual speech assumes a correlated role when there is redundancy between the information provided by vision and audition. As discussed above, the visible articulators that determine the vocal resonances, such as the lips, teeth and tongue, as well as ancillary movements, such as facial, head and hand movements, are temporally correlated with the vocalised acoustic signal (Chandrasekaran et al., 2009, Munhall et al., 2004a, Yehia et al., 2002, Grant and Seitz, 2000). Thus, in noisy environments, these visual cues provide important information pertaining to the timing of the target signal. Based on the “attention in time” hypothesis (Large and Jones, 1999, Jones et al., 2006, Nobre et al., 2007, Nobre and Coull, 2010), this would allow the listener to increase their attentional allocation at the correct moments in time, helping them direct auditory analysis to the speech signal of interest, rather than the surrounding background noise (Summerfield, 1987, Peelle and Sommers, 2015, Grant and Seitz, 2000). Tuning into the temporal pattern of continuous speech may also help listeners know when to expect certain types of acoustic speech information, thus helping them decode segmental (e.g., phonemes, syllables, words) and suprasegmental (e.g., intonation, stress, rhythm) categories (Peelle and Sommers, 2015, Peelle and Davis, 2012, Summerfield, 1987). It has also been suggested that redundancies in AV speech could lead to a reduction in the allocation of cognitive resources, enhancing cognitive function overall and as a result improving speech comprehension (Alais et al., 2010).

In addition to providing temporal cues relating to the acoustic signal, visual speech also conveys information about the place and manner of articulation. In circumstances where acoustic information is ambiguous or degraded, visual cues such as the configuration of the vocal tract, mouth opening and closure, mouth shape, as well as configurations of the lips, teeth and tongue can provide complementary information (Campbell, 2008). For example, place of articulation provides critical distinctions between certain consonants such as /d/ and /p/ (Peelle and Sommers, 2015). Given that this information is distinguished acoustically by differences in F2 formant space, it is highly susceptible to masking in noisy environments (Miller and Nicely, 1955) and in those with impaired hearing (Walden et al., 1975). Thus, the availability of place of articulation in the visual signal provides a complementary source of information that allows us to distinguish between words such as ‘mad’ and ‘map’ (Peelle and Sommers, 2015). Manner of articulation is also sometimes visible; in the previous example for instance, the /p/ in ‘map’ can produce a visible lip-puff, which is absent when ‘mad’ is uttered (Campbell, 2008).

The phoneme is considered the smallest unit of acoustic speech that distinguishes one word from another in any language (Chomsky and Halle, 1968). The smallest unit of visual speech, the viseme, refers to speech gestures that are commonly confused during visual-only speech (Fisher, 1968, Miller and Nicely, 1955). However, the mapping from phonemes to visemes is not one-to-one; rather, multiple phonemes can map to a single viseme, meaning that visual speech generally presents more ambiguity than auditory speech. For example, ‘cap’ and ‘cab’ are readily distinguished acoustically by the voiced nature of /b/ relative to the unvoiced /p/, whereas visually, these words are almost identical. While visual speech does not offer complimentary information for every phoneme, in many cases it can help disambiguate acoustically confusable phonemes, in addition to providing robust temporal cues (Peelle and Sommers, 2015). In order to exploit correlated and complementary speech cues, the brain must integrate this information together. The next section describes AV speech integration in the brain from a behavioural, neurophysiological and theoretical perspective.

2.4 Multisensory Integration in AV Speech Processing

Multisensory integration is the process by which our brain combines information from two or more sensory modalities in order to enhance our perception of the world (Stein et al., 2014, Stein and Stanford, 2008). Given that speech (which is biologically significant to humans) is naturally a multisensory event, it is unsurprising that our brains exploit the correlations and redundancies in these signals to maximise the likelihood of us understanding our interlocutor.

Most of our understanding of how the brain integrates AV information comes from electrophysiology in the cat superior colliculus (SC; Stein and Meredith, 1993), a subcortical structure common to all mammalian brains. Certain neurons in SC were shown to respond to both auditory and visual stimulation, but responded in a non-linear manner to simultaneous AV stimuli (Fig. 2.7; Meredith and Stein, 1983, Meredith and Stein, 1985). Subsequent work yielded three fundamental principles of multisensory integration (Meredith and Stein, 1986a, Meredith et al., 1987): the interaction effect was largest when the signals occur at the same location (spatial rule), at the same time (temporal rule) and when the signals are minimally effective (principle of ‘inverse effectiveness’).

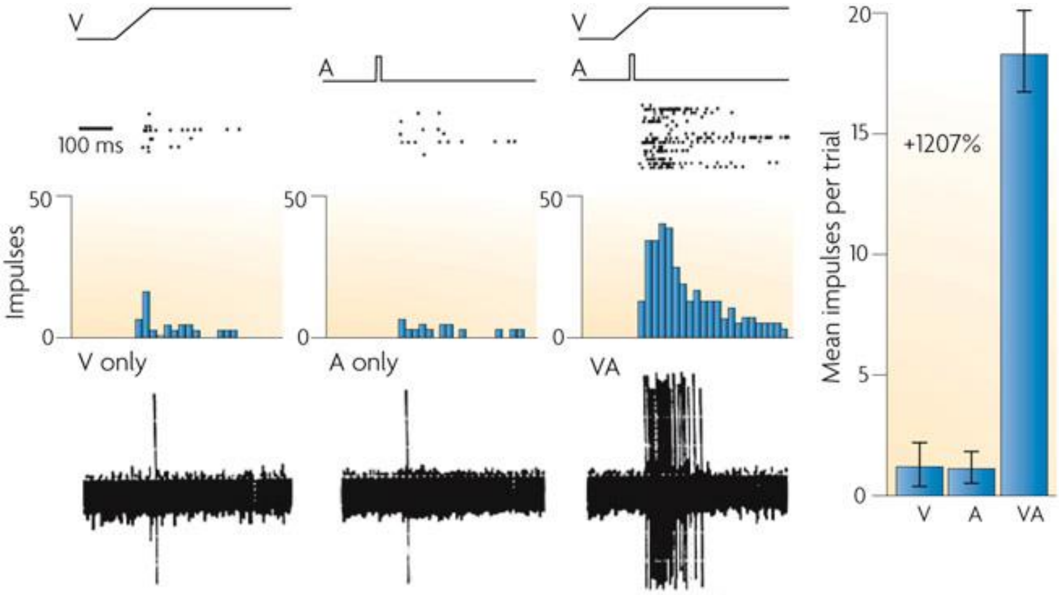


Figure 2.7: Multisensory integration in a superior colliculus neuron (adapted from Stein and Stanford, 2008).

2.4.1 Quantifying Multisensory Integration

One of the challenges in studying multisensory integration is how to isolate and quantity contributions from multisensory interactions. There have been numerous models developed to quantify multisensory integration based on neurophysiological and behavioural data (reviewed in Stevenson et al., 2014a). Most of these models assess multisensory integration based on two simple criteria: the maximum criterion or the additive criterion (Fig. 2.8*A*; Peelle and Sommers, 2015).

The maximum criterion model compares the response to a multisensory stimulus with that of the most effective unisensory condition (Meredith and Stein, 1983, Meredith and Stein, 1986b). The rationale is that any response measure departing from that of the most effective unisensory condition should be attributed to the multisensory nature of the stimulus, that is, to interactions between the inputs from the two modalities. When measuring behaviour, this model is only suitable when performance is either below threshold or near ceiling in at least one of the unisensory conditions (Stevenson et al., 2014a). In neurophysiology, this model can be applied when the signal being recorded is from a site that is only particularly responsive to unisensory stimulation from one modality, but displays enhanced responsiveness during multisensory stimulation. The maximum criterion model defines multisensory integration (MSI) as follows:

$$\mathrm{MSI} = AV - \max(A,V), \tag{2.1}$$

where variables *A*, *V*, and *AV* represent the behavioural/neurophysiological measures (e.g., accuracy, spike rate, amplitude) for each stimulus condition. Positive MSI values indicate enhancement, negative values indicate reduction and zero values indicate no integration (Fig. 2.8B; Meredith and Stein, 1983, Peelle and Sommers, 2015). Of course, when examining something like reaction time (RT), this model can be modified to compare *AV* with min(*A*,*V*), i.e., the fastest unisensory condition.

The additive criterion model on the other hand, compares the response to a multisensory stimulus with that of the algebraic sum of the unisensory conditions (Stein and Meredith, 1993, Barth et al., 1995, Berman, 1961). The rationale here is that the response to a multisensory stimulus should be equal to the sum of the responses generated separately by the two unisensory stimuli, if the two unisensory signals were processed independently. Thus, any departure from the summed response should be attributed to multisensory interactions (Besle et al., 2004b). For behavioural measures, this model is most suitable when the unisensory response magnitudes from both modalities are not near threshold or ceiling (Stevenson et al., 2014a). In neurophysiology, this approach is most suited to recording sites that are responsive to both unisensory stimuli, particularly when recording from populations of neurons. Based on the additive criterion, multisensory integration is defined as follows:

$$\mathrm{MSI} = AV - (A+V). \tag{2.2}$$

Here, positive MSI values indicate ‘superadditivity’, negative values indicate ‘subadditivity’ and zero values indicate no integration (Fig. 2.8C; Stein and Meredith, 1993, Peelle and Sommers, 2015). The validity of the additive model is well established, particularly in the field of electrophysiology (Besle et al., 2004b). This is because when measuring electric signals elicited by the brain, their magnitude is governed by the law of superposition of electric fields. The principle of superposition states that the net response of a linear system (and tissue is a linear conductor at macroscopic scales) at a given position and time caused by two or more stimuli is equal to the sum of the responses which would have been produced by each stimulus individually.

However, behavioural measurements are sometimes represented as probabilities (e.g., detection accuracy, RT), meaning it is necessary to include an expression of the joint unisensory probability in the model. For instance, if detection accuracy was being measured, this would account for the probability that a multisensory stimulus was detected in both modalities (Stevenson et al., 2014a). Or if RT was being measured, this would account for the probability that the stimuli were detected at the same time in both modalities. Suppose each variable in Eq. 2.2 represented the probability of detecting a stimulus in each condition, the formula could be extended to account for joint unisensory probability as follows:

$$\mathrm{MSI} = AV - (A + V - A \times V). \tag{2.3}$$

This is equivalent to assuming that an error in the AV condition only occurs if there is an incorrect response in both of the unisensory conditions, i.e., $1-AV = (1-A)\ (1-V)$ (Blamey et al., 1989). The same model can also be applied to RT measurements by replacing each variable in Eq. 2.3 with the RT cumulative distribution function (CDF) for each condition. This is equivalent to sampling simultaneously from the unisensory RT distributions, taking the faster of the two unisensory RTs and then computing the CDF, i.e., the ‘race model’ (Raab, 1962). Violation of the race model (i.e., positive MSI values) indicates multisensory interactions or ‘co-activation’ (Miller, 1982, Molholm et al., 2002).

To quantify MSI in terms of percentage gain, Meredith and Stein (1983) defined an ‘interactive index’ that scaled MSI relative to the magnitude of their model:

$$\mathrm{Gain} = \frac{\mathrm{MSI}}{P} \times 100, \tag{2.4}$$

where P is the multisensory response predicted by the unisensory response values, i.e., max(A,V) or [A+V] or [A+V−A×V]. In other words, this represents the percentage gain in processing attributable to multisensory interactions relative to independent unisensory processing.

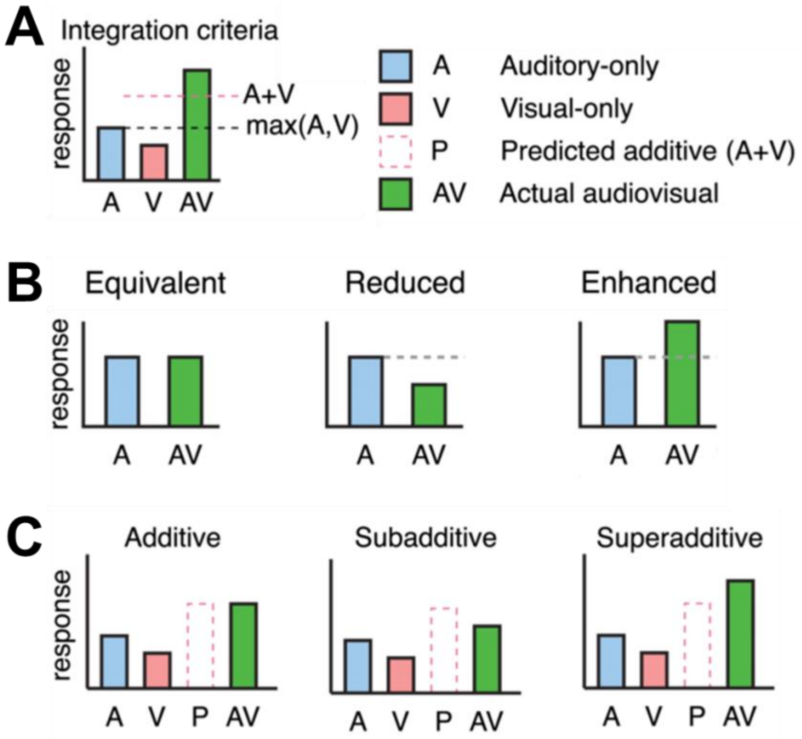


Figure 2.8: Quantification of multisensory integration.

A, Integration criteria. ***B***, Multisensory integration based on a maximum criterion model. ***C***, Multisensory integration based on an additive criterion model (adapted from Peelle and Sommers, 2015).

2.4.2 Behavioural Correlates

That visual speech could enhance our perception of auditory speech was first demonstrated behaviourally over 60 years ago (Sumby and Pollack, 1954, O'Neill, 1954). Specifically, it was shown that intelligibility was enhanced in noise, equivalent to an increase of up to 15 dB in signal-to-noise ratio (SNR; Sumby and Pollack, 1954), with a 1-dB improvement in SNR leading to a 5–10% increase in intelligibility, depending on speech materials (Miller et al., 1951). This led to the impression that visual speech only enhanced hearing in suboptimal listening conditions and (in line with the principle of inverse effectiveness) that this effect was inversely related to SNR and hearing ability (Erber, 1969, Erber, 1975, Erber, 1971, McCormick, 1979, Neely, 1956, Binnie et al., 1974). However, a re-examination of Sumby and Pollack's findings (Remez, 2005) showed that the benefit of AV speech is not limited to degraded acoustic environments. It has also been shown using extended passages of natural speech (instead of discrete tokens) that AV speech is beneficial in easy-to-hear (but hard-to-understand) environments, increasing the speed at which participants could repeat words in real time (Reisberg et al., 1987) as well as improving comprehension (Arnold and Hill, 2001). The former finding fits with studies that have observed faster RTs in response to AV syllables (Besle et al., 2004a, Klucharev et al., 2003). Note that enhanced intelligibility has been demonstrated at every level of speech, including syllables (Bernstein et al., 2004b), words (Sumby and Pollack, 1954) and sentences (Grant and Seitz, 2000).

It was recently argued (Ross et al., 2007a) that many of the early behavioural studies that demonstrated an inverse relationship between AV enhancement and SNR (i.e., inverse effectiveness) may have oversimplified this assumption. Several of these studies used a delimited set of word stimuli that were presented to the participants beforehand in the form of checklists (Sumby and Pollack, 1954, Erber, 1969, Erber, 1975). Thus, it is likely that speech-reading scores were artificially high due to familiarity, particularly at lower SNRs where intelligibility is more susceptible to ceiling effects (Ross et al., 2007a, Holmes, 2009, Bernstein et al., 2004a). Furthermore, measures of multisensory gain can be erroneously high depending on how it is calculated, i.e., absolute value versus relative percentage (Ross et al., 2007a, Holmes, 2009), and also by the density of the lexical neighbourhood of the word stimuli (Tye-Murray et al., 2007). To circumvent these shortcomings, Ross et al. (2007a) conducted a word recognition task (as opposed to detection) at multiple SNRs between 0 dB and auditory threshold (−24 dB). They presented a much larger set of words so that each presentation was unique and there were no checklists available to participants, which greatly reduced speech-reading accuracy (< 10%). In doing so, they demonstrated two very important behavioural aspects of AV speech: (1) the enhancement conferred by AV speech is far greater than that accounted for by speech-reading ability, i.e., it reflects multisensory interactions and (2) AV gain is greatest at an intermediate SNR (−12 dB), not at threshold (−24 dB). In other words, AV gain does not follow the principle of inverse effectiveness beyond a certain SNR.

Aside from studying how AV speech integration is impacted by background noise, many studies have investigated the impact of the sematic congruency between the auditory and visual streams. Much of this work was inspired by an influential study that accidentally demonstrated an interesting AV speech illusion, known as the McGurk effect (McGurk and MacDonald, 1976). The McGurk effect is a phenomenon whereby a particular incongruent pairing of auditory and visual syllables can produce the perceptual illusion of a syllable that was neither heard nor seen. For example, they found that when an auditory /ba/ was dubbed to a visual /ga/, the syllable perceived by participants was consistently /da/. The effect has since been replicated by numerous studies and in numerous different languages (Jiang and Bernstein, 2011, Summerfield and McGrath, 1984, Green and Kuhl, 1989, Sekiyama and Tohkura, 1991, Massaro et al., 1995). The McGurk illusion is even insensitive to knowledge of its basis (Campbell, 2008), although the nature of the fusion effect has been shown to be subject-dependent (Schwartz, 2010). Other than influencing how we perceive speech, incongruent AV pairings can alter our performance, delaying RTs relative to congruent AV speech and unimodal speech (Klucharev et al., 2003). While the McGurk effect has advanced out understanding of how the human brain integrates AV speech, it is usually perceived in a controlled experimental setting with well-synchronised AV stimuli and is not an illusion typically encountered in everyday life. It has been suggested that the spatial and temporal coherence of such incongruent stimuli may be a strong cue to their co-processing and 'binding' (Campbell, 2008).

The visual component of speech that contributes towards enhanced behaviour is not limited to just the mouth, but also movements of the head and eyebrows (Yehia et al., 2002, Munhall et al., 2004a, Thomas and Jordan, 2004), and even haptic movements (Fowler and Dekle, 1991). Furthermore, it has been shown that humans typically perceive desynchronised AV syllables as occurring simultaneously for audio leads of up to 90 ms and audio lags of up to 170 ms, and perceive McGurk fusion effects for audio leads of up to 30 ms and audio lags of up to 170 ms (Fig. 2.9; van Wassenhove et al., 2007, Miller and D'Esposito, 2005, Grant et al., 2004). This ~250 ms window of integration corresponds roughly to the average length of a syllable, thus it has been suggested that syllables may be an important unit of computation in AV speech processing (van Wassenhove, 2013). Furthermore, it has been shown that this window is narrower and more asymmetric for speech versus non-speech stimuli, in support of the notion that this tolerance is fine-tuned to the natural statistics of AV speech (Maier et al., 2011). It has also been suggested that the brain tolerates AV asynchronies because of the differences in the speeds of sight and sound, as well as differences in transduction times and neural latencies (Alais et al., 2010). Although asynchrony detection has not been shown to reflect speech reading ability (Grant and Seitz, 1998), it has been shown to predict susceptibility to the McGurk effect (Stevenson et al., 2012b).

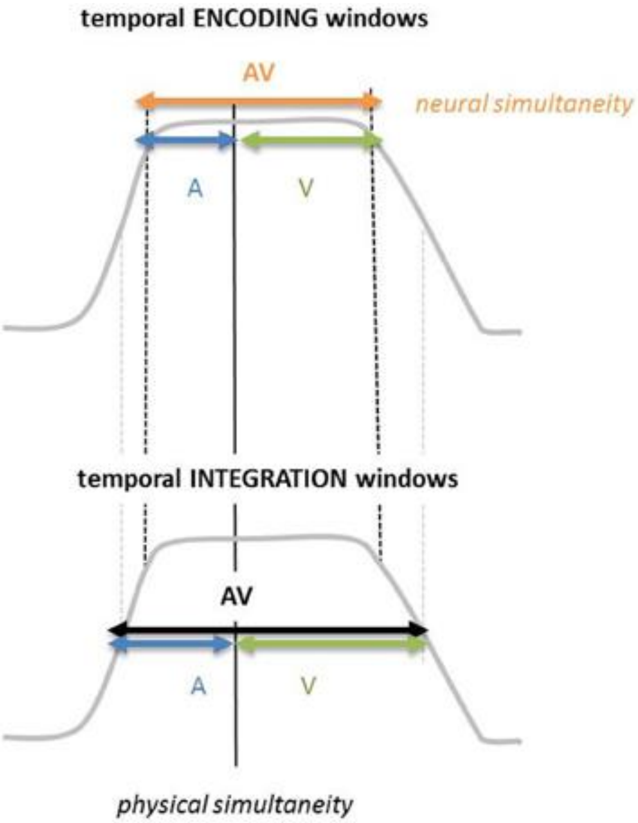


Figure 2.9: Temporal window of encoding and integration in AV speech.
The temporal encoding window (top) represents the time necessary for speech encoding and the temporal integration window (bottom) represents the encoding window plus the tolerated temporal noise leading to suboptimal encoding performance (Adapted from van Wassenhove, 2013).

2.4.3 Cortical Brain Regions in AV Speech Processing

A number of cortical brain regions have been linked to multisensory integration, none more so than the posterior STS (pSTS; Peelle and Sommers, 2015, Campbell, 2008, Alais et al., 2010). Numerous fMRI studies have implicated the pSTS as a primary binding area for multisensory speech processing as it is consistently activated during unisensory speech (both auditory and visual) as well as multisensory speech (Calvert et al., 1997, Callan et al., 2004, Wright et al., 2003, Arnal et al., 2009, Capek et al., 2004). It has been demonstrated that speech-reading tends to elicit both bilateral and left-lateralised activation (Capek et al., 2004, Calvert and Lewis, 2004, Bernstein et al., 2008). However, the nature of this activation is quite variable; left pSTS can (but does not always) exhibit superadditive activation during congruent AV speech (Calvert et al., 2000, Wright et al., 2003, Miller and D'Esposito, 2005), while subadditive activation has been reported in other superior temporal regions (Wright et al., 2003), and for incongruent AV speech in posterior STS (Calvert et al., 2000). Importantly, multisensory activation of pSTS has been demonstrated using both discrete speech stimuli (Miller and D'Esposito, 2005, Wright et al., 2003) and extended speech passages of natural speech (Yi et al., 2014, Calvert et al., 2000). A recent review highlighted that most reports of multisensory interactions in cortical areas are statistically weaker than superadditive (Alais et al., 2010). However, similar to the inverse effectiveness observed in SC neurons, STS activation becomes more superadditive at lower SNRs (Stevenson and James, 2009).

Other cortical brain regions have been identified as being putatively multisensory. Cortical recordings in the cat have shown the ectosylvian sulcus to contain multisensory neurons. These neurons exhibit superadditive responses to spatiotemporally coherent multisensory inputs, as well as inverse effectiveness and depression in response to disparate inputs (Stein and Wallace, 1996). In primates, multisensory neurons are commonly identified in *posterior parietal cortex* (PPC). While non-linear responses in PPC also require multisensory inputs to be spatiotemporally coincident, they have been shown to exhibit both superadditivity and subadditivity to such stimuli (Avillac et al., 2005). As well as caring about spatiotemporal coherence, it is likely that many cortical areas also care about the semantic congruency between multisensory inputs (Alais et al., 2010). In support of this notion, single-unit recordings in macaque STS suggest that sematic congruency is necessary to elicit multisensory interactions in certain neurons (Barraclough et al., 2005, Beauchamp et al., 2004). While STS is generally considered the "*sine qua non*" of AV speech integration, other studies have demonstrated greater activation in areas such as the left supramarginal and angular gyrus (Bernstein et al., 2008).

In addition to such putatively multisensory areas, single-unit recordings in primates have revealed early AV interactions in auditory cortex (Ghazanfar et al., 2005, Kayser et al., 2010, Kayser et al., 2008, Chandrasekaran et al., 2013). So too have intracranial recordings in humans (Besle et al., 2008, Mercier et al., 2015). fMRI research has also demonstrated AV interactions in auditory cortex in both primates (Kayser et al., 2007, Kayser et al., 2009) and humans (Okada et al., 2013). This fits with previous fMRI studies that have shown activation in human primary auditory cortex during silent lipreading (Pekkola et al., 2005, Calvert et al., 1997). In addition to sensory cortical areas, speech reading has also been shown to activate motor regions such as Broca’s area (BA 44/45) and anterior parts of the insula (Campbell et al., 2001, Watkins et al., 2003, Ojanen et al., 2005, Skipper et al., 2005, Fridriksson et al., 2008). It has been suggested that observing a speaker’s articulatory movements may activate motor plans than in turn preferentially access phonemic categories, shaping the listener’s auditory perceptual experience (Peelle and Sommers, 2015, Tian and Poeppel, 2012, Möttönen and Watkins, 2009).

2.4.4 A Multistage Model of AV Speech Processing

Before there was sufficient evidence to suggest that AV speech was integrated in sensory-specific regions, it was thought that auditory and visual information were initially processed independently over successive stages within their respective unisensory pathways and combined later in higher-order, association areas (Massaro, 1999, Grant et al., 1998). Recent perspectives on multisensory integration have redefined the role of primary sensory areas, suggesting that crossmodal information can influence processing in such regions at an early stage (Foxe and Schroeder, 2005, Schroeder and Foxe, 2005, Driver and Noesselt, 2008, Kayser and Logothetis, 2007, Ghazanfar and Schroeder, 2006). Acceptance of both early and late integration models has led to the conception of a new ‘multistage’ integration model (Peelle and Sommers, 2015).

Early and late integration models have been linked to different brain regions, thus, it is likely that each stage is underpinned by different neural mechanisms and serves a complementary role in AV speech perception. It has been suggested that early integration serves to increase the sensitivity of auditory cortex to incoming acoustic information (Peelle and Sommers, 2015). This theory is supported by the aforementioned electrophysiological studies that have observed early AV interactions in primate auditory cortex (Ghazanfar et al., 2005, Kayser et al., 2010, Lakatos et al., 2007, Kayser et al., 2008). This is also supported by the fact that visual speech cues reliably precede and predict auditory speech information (Schwartz and Savariaux, 2014, Chandrasekaran et al., 2009). Furthermore, such natural asynchronies in AV communication have been shown to regulate and enhance multisensory interactions in auditory cortex (Kayser et al., 2008, Perrodin et al., 2015). Preceding visual information could be projected to auditory cortex in several different ways (Fig. 2.10; Peelle and Sommers, 2015): directly in a thalamocortical feedforward manner (Foxe and

Schroeder, 2005, Besle et al., 2008), laterally in a corticocortical manner (Falchier et al., 2002, Schroeder et al., 2008, Arnal et al., 2009) or indirectly via supramodal regions such as pSTS (Ghazanfar et al., 2005) or frontal regions such as BA 44/45, which has specifically been linked to speech-reading (Campbell et al., 2001, Skipper et al., 2005, Watkins et al., 2003, Ojanen et al., 2005). Thereafter, visual cues could increase the sensitivity of auditory cortex via crossmodal phase-resetting, whereby relatively discrete visual landmarks reset the phase of ongoing low-frequency oscillations in auditory cortex, such that the arrival of upcoming acoustic information coincides with a high excitability phase of the auditory neuronal population (Schroeder et al., 2008, Lakatos et al., 2007, Kayser et al., 2008). The efficacy of such a mechanism in the context of continuous speech has been linked to the relationship between the temporal scale of segmental/suprasegmental speech units and the hierarchically coupled rhythmic oscillatory complex in auditory cortex (Schroeder et al., 2008).

It is believed that late integration serves to constrain the possible candidates in a spoken utterance based on visual information about a speaker's articulators (Peelle and Sommers, 2015). Late integration models are generally associated with higher-order ‘multisensory’ areas such as those mentioned in section 2.2.3. Such areas could provide a cortical site where feedforward auditory and visual information could converge and become ‘bound’ to from a multisensory object, as well as providing feedback to unimodal areas (Driver and Spence, 2000). The STS, which responds to both A and V inputs, has been suggested as a likely candidate for a binding site and a feedback provider to auditory cortex (Calvert and Campbell, 2003, Beauchamp et al., 2004, Arnal et al., 2009, Kayser and Logothetis, 2009). Mechanistically, binding of auditory and visual speech features could depend on their temporal coherence (Senkowski et al., 2008, Maier et al., 2008). In auditory scene analysis, it has been hypothesised that multi-feature auditory sources are segregated into perceptual streams based on the temporal coherence of their acoustic features (Elhilali et al., 2009, Shamma et al., 2011). Similarly, visual speech cues, being correlated with the spectrotemporal dynamics of the auditory speech signal, could result in visual features being bound to the auditory features to form a multisensory object. This notion fits within an ‘analysis-by-synthesis’ framework of speech processing, which proposes that speech is first analysed by breaking it up into its constituent spectrotemporal channels and that auditory objects (or multisensory objects) are synthesised from those channels that modulate together in a temporally coherent manner (Ding et al., 2014, Ding and Simon, 2014). As well as projecting forward to higher-order areas, STS could also feedback to unisensory cortex. It has been suggested that this process could serve to differentially weight auditory and visual information depending on how informative or reliable each modality is (Peelle and Sommers, 2015). In support of this, Nath and Beauchamp (2011) demonstrated that differentially varying the reliability of the auditory and visual streams in AV speech stimuli directly affected the functional connectivity between STS and the corresponding unisensory cortex. However, another study showed that increasing viseme ambiguity increased functional connectivity between STS and motion-sensitive visual areas, as well as increasing connectivity between STS and auditory cortex (Arnal et al., 2009).

In summary, a multistage integration model accounts for both early AV interactions in auditory cortex, and a later integration stage that binds crossmodal features based on their temporal coherence, as well as differentially weighting each modality based on its reliability. Multistage models somewhat explain the difficulty in predicting AV speech performance based on a single measure of integration because single-stage models cannot account for interactions between multiple integration mechanisms (Peelle and Sommers, 2015).

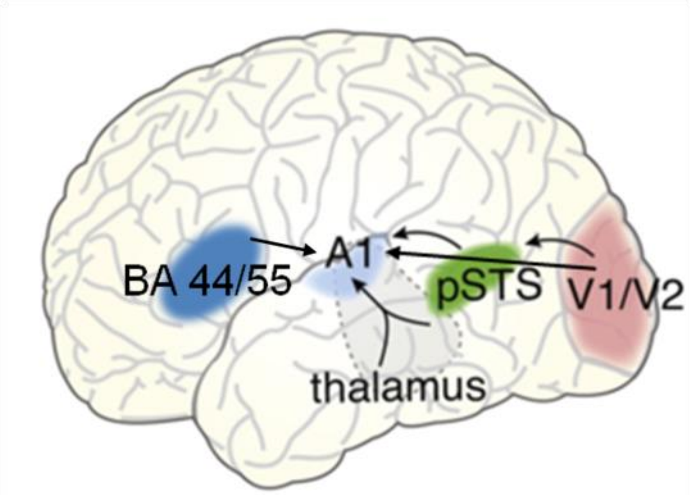


Figure 2.10: Anatomical pathways for routing visual information to auditory cortex (adapted from Peelle and Sommers, 2015).

2.5 An EEG/MEG Account of AV Speech Processing

Much of our knowledge on the temporal dynamics of AV speech processing comes from single-unit recordings in the primate brain. While many primate studies have used naturalistic stimuli such as monkey vocalisations in order to predict how humans process and integrate multisensory speech (Ghazanfar et al., 2005, Chandrasekaran et al., 2013, Kayser et al., 2008, Kayser et al., 2010), it is not directly comparable to human speech as it lacks the lexical complexity of human conversation. Intracranial (ECoG) recordings have been used to study AV speech in humans (Besle et al., 2008, Mercier et al., 2015), however, this highly invasive technique can only be conducted in pre-surgical epileptic patients, hence its literature is limited.

EEG, and to a lesser extent MEG, have been used to examine the timecourse of AV speech processing, mostly in response to discrete speech tokens such as syllables and words rather than natural continuous speech (but see Zion-Golumbic et al., 2013a, Luo et al., 2010). The reason for this is partly methodological; a reliable technique has not yet been developed for quantifying a neural measure of multisensory integration in the context of continuous AV speech. The following section is devoted to a review of the ERP/ERF-based AV speech literature and is followed by a discussion on the current techniques used for studying continuous speech.

2.5.1 Measuring AV Integration with Evoked Responses

Much of the EEG/MEG-based literature on AV speech integration has focused on how cortical activity responds to discrete, time-locked AV syllables. While MEG can approximate the cortical response from localised sources (e.g., auditory cortex, STS), EEG is generally studied in sensor space (i.e., at the scalp level), making it more difficult to accurately localise AV interaction effects. EEG responses (i.e., ERPs) are typically measured at temporal, centro-parietal and fronto-central scalp locations. The most common finding amongst EEG studies is that certain ERP components (typically the N1/P2 complex) are significantly reduced in amplitude by AV presentation compared to audio-only presentation or additive models (Besle et al., 2004a, van Wassenhove et al., 2005, Klucharev et al., 2003, Stekelenburg and Vroomen, 2007, Pilling, 2009). Similarly, MEG studies have observed reduced amplitude in the M100 ERF component in response to AV syllables (Arnal et al., 2009). It has also been shown that such AV interactions occur earlier in auditory cortex compared to those in STS (Möttönen et al., 2004), suggesting that they index non-phonetic and phonetic AV interactions respectively (Klucharev et al., 2003). The majority of the aforementioned studies report only late (>100 ms) AV interactions (Bernstein et al., 2008, Besle et al., 2004a, Möttönen et al., 2004, Stekelenburg and Vroomen, 2007, van Wassenhove et al., 2005), however this may be a result of EEG/MEG's insensitivity to early- and middle-latency response components (Picton, 2013).

Interestingly, incongruent or McGurk AV pairings have also been shown to elicit a similar reduction in ERP/ERF amplitude (Stekelenburg and Vroomen, 2007, van Wassenhove et al., 2005, Arnal et al., 2009, Klucharev et al., 2003), further demonstrating the autonomous nature of multisensory integration in the context of an artificially created McGurk scenario. In support of claims by Campbell (2008), that such autonomous integration is mediated by the temporal congruency of McGurk stimuli, it was demonstrated that asynchronous AV speech (where A leads V) does not elicit a reduction in amplitude (Pilling, 2009). Furthermore, another study showed that such interactions only occur if the visual stream precedes the acoustic stream, as is typically the case in natural speech (Stekelenburg and Vroomen, 2007). Despite this, visual predictability does not modulate the degree of amplitude reduction (van Wassenhove et al., 2005, Arnal et al., 2009). It has been suggested that ERP/ERF amplitude reduction could be underpinned by a deactivation mechanism that minimises the processing of redundant cross-modal information based on an internal prediction derived from the preceding visual input (van Wassenhove et al., 2005, Arnal et al., 2009), an idea known as *predictive coding* (Friston, 2005). This theory also fits well with the idea of a late integration mechanism that constrains lexical selection (Peelle and Sommers, 2015).

While the aforementioned studies have mainly focused on multisensory interactions in auditory ERPs, a more recent study examined the effects of integration on visual ERPs for various levels of visual salience and demonstrated some interesting effects (Stevenson et al., 2012a). Specifically, they showed that the P1-N1 complex became more subadditive as salience decreased and that the N1-P2 complex became less subadditive as salience decreased. While they interpreted the latter as evidence for inverse effectiveness, in light of the above literature which predominantly identified AV interactions by a reduction in amplitude, it could be argued that the former result was instead reflective of inverse effectiveness and not the latter. However, they found that the multisensory gain across sensory levels in the N1-P2 complex, i.e., $\Delta\mathrm{AV}-\left(\Delta\mathrm{A}+\Delta\mathrm{V}\right)$, was positively (but not significantly) correlated with the multisensory gain in RT (as measured by violation of the race model). A neural index of multisensory gain that reliably predicts a behavioural metric of multisensory gain is something that has not yet been established.

Another important effect reported in many of these studies is that of a multisensory latency effect. Some of the studies (but not all of them) that reported attenuation of ERP/ERF components during AV speech also noted a marked facilitation in the latency of the same components (van Wassenhove et al., 2005, Arnal et al., 2009, Stekelenburg and Vroomen, 2007). Specifically, they found that the attenuated response components peaked at an earlier latency relative to audio-only/additive responses. However, these “shifted” components appeared to onset at the same latency and with the same slope as those elicited by audio-only speech, except that they peaked earlier. Thus, it has been suggested that this earlier peak in amplitude may be a direct consequence of the reduction in amplitude, i.e., an artefact (Stekelenburg and Vroomen, 2007). This can easily be explained by the laws of superposition which govern the summation of ERP responses (Handy, 2005). However, this latency effect fits well with the fact that visual speech naturally precedes and predicts auditory speech dynamics (Chandrasekaran et al., 2009, Schwartz and Savariaux, 2014). Furthermore, it was found that the magnitude of this latency shift was proportional to the degree to which the visual stimulus predicted the auditory stimulus (van Wassenhove et al., 2005, Arnal et al., 2009). This also fits with animal models which have shown that neurons in macaque auditory cortex fire earlier in response to AV vocalisations (Chandrasekaran et al., 2013), as do neurons in the cat SC in response to AV stimuli (Rowland et al., 2007).

2.5.2 Studying AV Integration Using Continuous Speech

While ERP/ERF techniques have yielded great insight into the dynamic processing of multisensory speech, they have certain limitations which new approaches have sought to overcome. Firstly, there is the aforementioned confound of latency versus amplitude interactions. Disentangling this issue is problematic because there is no way to separate out contributions from different modalities. Secondly, they must be derived by averaging the response to discrete, isolated tokens that do not reflect natural, continuous speech. Lastly, ERPs/ERFs do not reflect the true temporal dynamics of the auditory system’s response because discrete stimuli must have some arbitrary duration which, assuming convolution, smears the response over time.

Indeed, a way to circumvent the issues involved in estimating the cortical response of the auditory system is to avoid estimating it altogether. One such approach that has become popular in recent years (particularly in MEG) is to measure the consistency of neural phase patterns over repeated trials of extended speech (Luo and Poeppel, 2007). The rational here is that the consistency in phase across repeated trials indexes the degree to which the cortical activity tracks the speech signal. This can be quantified using an inter-trial (phase) coherence measure (Luo and Poeppel, 2007, Howard and Poeppel, 2010, Peelle et al., 2013) or an inter-trial correlation measure (Ding and Simon, 2013, Ding et al., 2014). Some of these studies (but not all of them) have shown preferential tracking in theta band (4–7 Hz) activity and thus have implicated the syllable as a computational primitive in speech processing (Luo and Poeppel, 2007, Giraud and Poeppel, 2012, Howard and Poeppel, 2010, Peelle et al., 2013). However, this effect is likely attributable to the use of short (<5 s duration) single-sentence stimuli which do not contain much delta-frequency information below ~2 Hz, i.e., prosodic content (Ding et al., 2014). It has been shown using longer speech stimuli (>30 s in duration) that cortical activity reliably tracks the speech envelope below ~10 Hz and even more reliably below ~4 Hz, especially in noise (Ding and Simon, 2013, Ding et al., 2014). However, a recent perspective on speech tracking suggests that theta entrainment likely encodes speech-specific features, critical for intelligibility, while delta entrainment likely encodes non-speech-specific acoustic rhythm (Ding and Simon, 2014).

As well as being used to study how the human brain processes continuous auditory speech, this approach has been applied in AV speech studies using natural speech and revealed some important insights into the underlying neural mechanisms of AV integration (Zion-Golumbic et al., 2013a, Luo et al., 2010). Specifically, it has been demonstrated that auditory cortex tracks both auditory and visual stimulus dynamics in delta and theta band responses (Luo et al., 2010). This provides evidence in support of the theory that the phase of auditory cortical activity could be reset by ongoing phasic variations in visual cortex (see section 2.4.4). In another study that examined AV speech in a cocktail party scenario, it was shown that delta and theta band responses in auditory cortex tracked the attended speech signal more reliably during AV presentation, compared to audio-only presentation (Zion-Golumbic et al., 2013a). However, they found that AV speech did not enhance cortical tracking when participants only listened to a single-speaker. They admit, however, that inter-trial coherence is an indirect measure of speech tracking and, as such, may be insensitive to subtle differences in more easy-to-hear environments.

This, coupled with the fact that such an approach does not allow for characterisation of the auditory system’s response, highlights the need to develop a new technique for studying how the brain processes natural, continuous AV speech stimuli. The next chapter describes one such approach that was developed as a MATLAB toolbox for the purpose of this research work.

Chapter 3 The Multivariate Temporal Response Function (mTRF) Toolbox: a MATLAB Toolbox for Relating Neural Signals to Continuous Stimuli

3.1 Introduction

As detailed in Chapter 2, much of the previous research on the electrophysiology of AV speech processing has focused on the rather special case of isolated, discrete speech stimuli (Besle et al., 2004a, van Wassenhove et al., 2005, Arnal et al., 2009, Klucharev et al., 2003, Stekelenburg and Vroomen, 2007, Pilling, 2009, Sams et al., 1991, Möttönen et al., 2004). While more recent studies have focused on how the phase of neural signals reflect the dynamics of ongoing speech (Zion-Golumbic et al., 2013a, Luo et al., 2010), the methodological approach involved does not facilitate characterisation of the auditory system's response, and in any case, is an indirect measure of speech tracking.

A more direct way to investigate neural speech tracking is to mathematically model a function that describes the way a particular property of the speech stimulus is mapped onto neural responses, i.e., the aforementioned technique of system identification (Marmarelis, 2004). While there are several classes of models that can be implemented for this purpose (reviewed in Wu et al., 2006), the most straightforward class are linear time-invariant (LTI) systems. Although the human brain is neither linear nor time-invariant, LTI systems can be completely characterised by their impulse response (see Fig. 3.1). An SI method known as *reverse correlation* has become a common technique for characterising LTI systems in neurophysiology (Ringach and

Shapley, 2004), an approach that has long been established in both visual and auditory animal neurophysiology (Coppola, 1979, Marmarelis and Marmarelis, 1978, De Boer and Kuyper, 1968). Analogous to ERP/ERF-based methods, this technique approximates the impulse response of the sensory system under investigation, except it does not require that the stimulus be discrete in time (e.g., Fig. 3.1, bottom). Moreover, it circumvents many of the aforementioned issues related to using ERPs/ERFs (see section 2.5.2). Reverse correlation in its simplest form can be implemented via a straightforward cross-correlation between the input and output of an LTI system (Ringach and Shapley, 2004). While this approach has been used to study how speech is encoded in human EEG/MEG (Ahissar et al., 2001, Abrams et al., 2008, Aiken and Picton, 2008), it is better suited to stimuli modulated by a stochastic process such as Gaussian white noise (see example in section 3.4.5). As such, most instances of this approach in animal models have traditionally used white noise stimuli (Coppola, 1979, Marmarelis and Marmarelis, 1978, De Boer and Kuyper, 1968, Eggermont et al., 1983, Ringach et al., 1997). This work has even inspired researchers to investigate how such stochastic signals are encoded in the human brain (Lalor et al., 2006, Lalor et al., 2009).

That said, the human brain has evolved to process ecologically relevant stimuli that rarely conform to a white random process. For example, in the context of human neuroscience research, a proper understanding of how the brain processes natural speech would surely require that natural speech be used as a stimulus in the laboratory, given that neurons respond differently to more complex stimuli (Theunissen et al., 2000). As such, researchers using animal models have shifted their focus towards studying the brain using more natural stimuli thanks to the development of SI methods such as *normalised reverse correlation* (NRC; Theunissen et al., 2001), *ridge regression* (Machens et al., 2004) and *boosting* (David et al., 2007). Each of these techniques converge on the same theoretical solution but use different priors and, critically, give an unbiased impulse response estimate for non-white stimuli. This has inspired researchers to characterise the *spectrotemporal receptive fields* of auditory cortical neurons in various animal models (Depireux et al., 2001, Tomita and Eggermont, 2005). As a result, researchers interested in how human speech is processed have begun to model response functions describing the linear mapping between properties of natural speech (such as the envelope or spectrogram) and population responses in both animals (Mesgarani et al., 2008, David et al., 2007) and humans (Lalor and Foxe, 2010, Ding and Simon, 2012b). There have been similar efforts to model response functions relating more natural visual stimulus properties to neural responses in humans (Gonçalves et al., 2014), again inspired by single-unit electrophysiology work (Jones and Palmer, 1987, David and Gallant, 2005).

Most of the aforementioned studies have modelled the stimulus-response mapping function in the forward direction (i.e., forward modelling). However, this mapping can also be modelled in the reverse direction (i.e., backward modelling), offering a complementary way to investigate how stimulus features are encoded in neural response measures. Unlike forward models, backward model parameters are not readily neurophysiologically interpretable (see Haufe et al., 2014), but can be used to reconstruct or decode stimulus features from the neural response, a method known as *stimulus reconstruction*. This approach has several advantages over forward modelling approaches, especially when recording from population responses using multi-channel systems such as EEG. Firstly, because reconstruction projects back to the stimulus domain, it does not require preselection of neural response channels (Mesgarani et al., 2009). In fact, inclusion of all response channels in the backward model is advantageous because the reconstruction method gives zero weight to irrelevant channels whilst allowing the model to capture additional variance using channels potentially excluded by feature selection approaches (Pasley et al., 2012). Secondly, population responses recorded at different channels tend to be highly correlated (especially in EEG) which can bias the model. However, this is no longer an issue because the inter-channel correlation is removed from the reconstruction model (see section 3.2.4; Mesgarani et al., 2009). Thirdly, stimulus features that are not explicitly encoded in the neural response may be inferred from correlated input features that are encoded. This prevents the model from allocating resources to the encoding of redundant stimulus information (Barlow, 1972). The stimulus reconstruction method has previously been used to study both the visual and auditory system in various animal models (Bialek et al., 1991, Stanley et al., 1999, Rieke et al., 1995). More recently, it has been adopted for studying speech processing in the human brain using intracranial and non-invasive electrophysiology (Pasley et al., 2012, Mesgarani et al., 2009, Ding and Simon, 2013, O'Sullivan et al., 2015).

While numerous research groups have begun to regularly use different forms of SI to study the neural processing of natural speech (Di Liberto et al., 2015, Ding and Simon, 2014, Martin et al., 2014, Zion-Golumbic et al., 2013b, Mesgarani and Chang, 2012), the approach has not yet been widely adopted throughout the neuroscience community because of the challenge associated with its implementation. The goal of the present chapter is to introduce a recently-developed SI toolbox that provides a straightforward and flexible implementation of the ridge regression approach (Lalor et al., 2006, Machens et al., 2004). I begin by summarizing the mathematics underlying the approach, continue by providing some concrete examples of how the toolbox can be used and conclude by discussing some of its applications and considerations. The work from this chapter has resulted in the publication of an open-source MATLAB toolbox, known as mTRF Toolbox (http://sourceforge.net/projects/aespa/), which has already been downloaded and used by several international labs. The content of this chapter is currently being prepared for publication in a scientific methods journal.

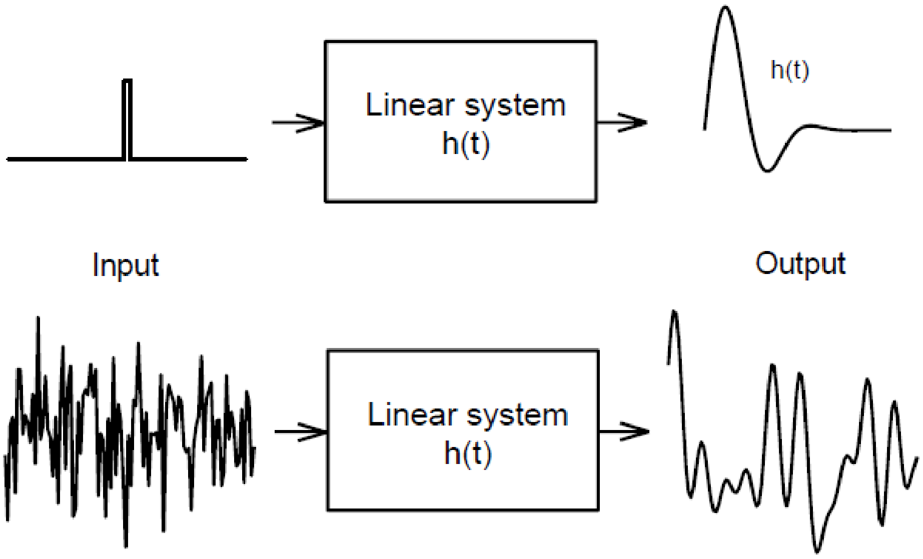


Figure 3.1: A linear time-invariant (LTI) system.
An LTI system can be characterised as the output to a discrete input (top), or by using a white noise input and cross-correlating the input and output (bottom; adapted from Ringach and Shapley, 2004).

3.2 The Ridge Regression Approach

3.2.1 Forward Models: Temporal Response Function Estimation

Forward models are sometimes referred to as generative or encoding models, because they describe how the system generates or encodes information (Haufe et al., 2014). Here, they will be referred to as temporal response functions (TRFs; Ding and Simon, 2012b). There are a number of ways of mathematically describing how the input to a system relates to its output. One commonly used approach – and the one that will be described in this chapter – is to assume that the output of the system is related to the input via a simple linear convolution. In the context of a sensory system where the output is monitored by N recording channels, let's assume that the instantaneous neural response, $r(t,n)$, sampled at times $t = 1\ldots T$ and at channel n consists of a convolution of the stimulus property, $s(t)$, with an unknown channel-specific TRF, $w(\tau,n)$. The response model can be represented in discrete time as:

$$r(t,n) = \sum_{\tau} w(\tau,n) s(t-\tau) + \varepsilon(t,n), \tag{3.1}$$

where $\varepsilon(t,n)$ is the residual response at each channel not explained by the model. Essentially, a TRF can be thought of as a filter that describes the linear transformation of the ongoing stimulus to the ongoing neural response. The TRF, $w(\tau,n)$, describes this transformation for a specified range of time lags τ relative to the instantaneous occurrence of the stimulus feature $s(t)$.

In the context of speech for example, $s(t)$ could be a measure of the speech envelope at each moment in time and $r(t,n)$ could be the corresponding EEG response at channel n. The range of time lags over which to calculate $w(\tau,n)$ might be that typically used to capture the cortical response components of an AEP, e.g., −100 to 400 ms. The resulting value of the TRF at −100 ms, would index the relationship between the speech envelope and the neural response 100 ms earlier (obviously this should have an amplitude of zero), whereas the TRF at 100 ms would index how a unit change in the amplitude of the speech envelope would affect the EEG 100 ms later (Lalor et al., 2009).

The TRF, $w(\tau,n)$, is estimated by minimizing the mean-squared error (MSE) between the actual neural response, $r(t,n)$, and that predicted by the convolution, $\hat{r}(t,n)$:

$$\min \varepsilon(t,n) = \sum_{t} \left[r(t,n) - \hat{r}(t,n) \right]^2. \tag{3.2}$$

In practice, this is solved using reverse correlation (De Boer and Kuyper, 1968), which can be easily implemented using the following matrix operations:

$$\mathbf{w} = \left(\mathbf{S}^{\mathrm{T}} \mathbf{S} \right)^{-1} \mathbf{S}^{\mathrm{T}} \mathbf{r}, \tag{3.3}$$

where **S** is the lagged time series of the stimulus property, **s**, and is defined as follows:

$$\mathbf{S}=\begin{bmatrix} s(1-\tau_{\min}) & s(-\tau_{\min}) & \cdots & s(1) & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots & s(1) & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots & \vdots & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \cdots & s(1) \\ s(T) & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ 0 & s(T) & \cdots & \vdots & \vdots & \cdots & \vdots \\ \vdots & 0 & \cdots & \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & s(T) & s(T-1) & \cdots & s(T-\tau_{\max}) \end{bmatrix}. \tag{3.4}$$

The values $\tau_{\min}$ and $\tau_{\max}$ represent the minimum and maximum time lags (in samples) respectively. In **S**, each time lag is arranged column-wise and non-zero lags are padded with zeros to ensure causality (Mesgarani et al., 2009). The window over which the TRF is calculated is defined as $\tau_{window} = \tau_{\max} - \tau_{\min}$ and the dimensions of **S** are thus $T \times \tau_{window}$. To include the constant term (y-intercept) in the regression model, a column of ones is concatenated to the left of **S**. In Eq. 3.3, variable **r** is a matrix containing all the neural response data, with channels arranged column-wise (i.e., a $T \times N$ matrix). The resulting TRF, **w**, is a $\tau_{window} \times N$ matrix with each column representing the univariate mapping from **s** to the neural response at each channel.

One of the important points here is that this analysis explicitly takes into account the autocovariance structure of the stimulus. In non-white stimuli, such as natural speech, the intensity of the acoustic signal modulates gradually over time, meaning it is correlated with itself at different time lags. A simple cross-correlation of a speech envelope and the corresponding neural response would result in temporal smearing of the impulse response function. A solution is to divide out the autocovariance structure of the stimulus from the model such that it removes the correlation between different time points. The TRF approach, which does this, is therefore less prone to temporal smearing than a simple cross-correlation approach (see section 3.4.5).

3.2.2 Regularisation

An important consideration when calculating the TRF is that of regularisation, i.e., introducing additional information to solve any *ill-posed problems* and prevent *overfitting*. The ill-posed problem has to do with inverting the autocovariance matrix, $\mathbf{S}^{\mathrm{T}}\mathbf{S}$. Matrix inversion is particularly prone to numerical instability when solved with finite precision. In other words, small changes in $\mathbf{S}^{\mathrm{T}}\mathbf{S}$ (such as rounding errors due to discretisation) could cause large changes in **w** if the former is ill-conditioned. This does not usually apply when the stimulus represents a stochastic process because $\mathbf{S}^{\mathrm{T}}\mathbf{S}$ would be full rank (Lalor et al., 2006). However, the autocorrelation properties of a non-white stimulus such as speech means that it is more likely to be singular (i.e., have a determinant of zero). Typically, numerical treatment of an ill-conditioned matrix involves reducing the variance of the estimate by adding a bias term or 'smoothing solution'. Addition of this smoothing term also solves the other main issue, that of overfitting. The reverse correlation analysis is utterly agnostic as to the biological nature of the data that it is being asked to model. As a result, without regularisation, the resulting TRF can display biologically implausible properties such as very high-frequency fluctuations. Regularisation serves to prevent overfitting to such high-frequency noise along the low-variance dimensions (Theunissen et al., 2001, Mesgarani et al., 2008).

In practice, both ill-posed problems and overfitting can be solved simultaneously by weighting the diagonal of $\mathbf{S}^{\mathrm{T}}\mathbf{S}$ before inversion, a method known as Tikhonov regularisation or ridge regression (Tikhonov and Arsenin, 1977):

$$\mathbf{w} = \left(\mathbf{S}^{\mathrm{T}}\mathbf{S} + \lambda\mathbf{I}\right)^{-1}\mathbf{S}^{\mathrm{T}}\mathbf{r}, \tag{3.5}$$

where **I** is the identity matrix and λ is the smoothing constant or 'ridge parameter'. The ridge parameter can be adjusted using cross-validation to maximise the correlation between $r(t,n)$, and $\hat{r}(t,n)$ (David and Gallant, 2005). TRF optimisation will be described in more detail in section 3.3.2. While this form of ridge regression enforces a smoothness constraint on the resulting model by penalising TRF values as a function of their distance from zero, another option is to quadratically penalise the difference between each two neighbouring terms of **w** (Lalor et al., 2006):

$$\mathbf{w} = \left(\mathbf{S}^{\mathrm{T}}\mathbf{S} + \lambda\mathbf{M}\right)^{-1}\mathbf{S}^{\mathrm{T}}\mathbf{r}, \quad \text{where}\,\mathbf{M} = \begin{bmatrix} 1 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & -1 & 2 & -1 \\ & & & & -1 & 1 \end{bmatrix}. \tag{3.6}$$

Tikhonov regularisation (Eq. 3.5) reduces overfitting by smoothing the TRF estimate in a way that is insensitive to the amplitude of the signal of interest. However, the quadratic approach (Eq. 3.6) reduces off-sample error whilst preserving signal amplitude (Lalor et al., 2006). As a result, this approach usually leads to an improved estimate of the system's response (as indexed by MSE) compared to Tikhonov regularisation.

3.2.3 Multivariate Analysis

The previous section focused on the specific case of relating a single, univariate input stimulus feature separately to each of multiple recording channels. However, most complex stimuli in nature are not processed as simple univariate features. When acoustic speech enters the ear, the signal is transformed into a spectrogram representation by the cochlea, consisting of multiple frequency bands which project along the auditory pathway (see section 2.1.2; Yang et al., 1992). The auditory system maps each of these frequency bands to the neural representation measured at the cortical level. This process can be modelled by a multivariate form of the TRF (i.e., mTRF).

Indeed, it is possible to define an mTRF that linearly maps a multivariate stimulus feature to each recording channel (Theunissen et al., 2000, Depireux et al., 2001). Using the above example, let $s(t, f)$ represent the spectrogram of a speech signal at frequency band $f = 1 \ldots F$. To derive the mTRF, the stimulus lag matrix, **S** (Eq. 3.4), is simply extended such that every column is replaced with F columns, each representing a different frequency band (i.e., a $T \times F\tau_{window}$ matrix). The resulting mTRF, $w(f, \tau, n)$, will be a $F\tau_{window} \times N$ matrix but can easily be unwrapped such that each independent variable is represented as a separate dimension (i.e., a $F \times \tau_{window} \times N$ matrix). Here, the constant term is included by concatenating F columns to the left of **S**.

An important consideration in multivariate TRF analysis is which method of regularisation to use. The quadratic regularisation term in Eq. 3.6 was designed to enforce a smoothness constraint and maintain SNR along the time dimension, but not any other. For high λ values, this approach would cause smearing across frequencies; hence it would not yield an accurate representation of the TRF in each frequency band. In this case, it will typically be most appropriate to use the identity matrix for regularisation (Eq. 3.5) so as to avoid enforcing a smoothness constraint across the non-time dimension of the mTRF – although, in some cases, this may actually be what is desired.

3.2.4 Backward Models: Stimulus Reconstruction

The previous sections describe how to forward model the linear mapping between the stimulus and the neural response. While this approach can be extended to accommodate multivariate stimulus features, it is suboptimal in the sense that it treats each response channel as an independent univariate feature. Backward modelling, on the other hand, derives a reverse stimulus-response mapping by exploiting all of the available neural data in a multivariate context. Backward models are sometimes referred to as discriminative or decoding models, because they attempt to reverse the data generating process by decoding the stimulus features from the neural response (Haufe et al., 2014). Here, they will simply be referred to as decoders.

Decoders can be modelled in much the same way as TRFs. Suppose the decoder, $g(\tau,n)$, represents the linear mapping from the neural response, $r(t,n)$, back to the stimulus, $s(t)$. This could be expressed in discrete time as:

$$\hat{s}(t)=\sum_{n}\sum_{\tau}r(t+\tau,n)g(\tau,n), \tag{3.7}$$

where $\hat{s}(t)$ is the reconstructed stimulus property. Here, the decoder integrates the neural response over a specified range of time lags τ. Ideally, these lags will capture the window of neural data that optimises reconstruction of the stimulus property. Typically, the most informative lags for reconstruction are commensurate with those used to capture the major components of a TRF, except in the reverse direction as the decoder effectively maps backwards in time. To reverse the lags used in the earlier TRF example ($\tau_{\min}=-100\,\text{ms}$, $\tau_{\max}=400\,\text{ms}$), the values of $\tau_{\min}$ and $\tau_{\max}$ are swapped but their signs remain unchanged, i.e., $\tau_{\min}=-400\,\text{ms}$, $\tau_{\max}=100\,\text{ms}$.

The decoder, $g(\tau,n)$, is estimated by minimizing the MSE between $s(t)$ and $\hat{s}(t)$:

$$\min\,\varepsilon(t)=\sum_{t}\left[s(t)-\hat{s}(t)\right]^2.$$

(3.8)

Analogous to the TRF approach, the decoder is computed using the following matrix operations:

$$\mathbf{g}=\left(\mathbf{R}^{\mathrm{T}}\mathbf{R}+\lambda\mathbf{I}\right)^{-1}\mathbf{R}^{\mathrm{T}}\mathbf{s} \tag{3.9}$$

where **R** is the lagged time series of the response matrix, **r**. For simplicity, **R** will be defined for a single-channel response system:

$$\mathbf{R} = \begin{bmatrix} r(1-\tau_{\min},1) & r(-\tau_{\min},1) & \cdots & r(1,1) & 0 & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots & r(1,1) & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots & \vdots & \cdots & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots & \cdots & r(1,1) \\ r(T,1) & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ 0 & r(T,1) & \cdots & \vdots & \vdots & \cdots & \vdots \\ \vdots & 0 & \cdots & \vdots & \vdots & \cdots & \vdots \\ \vdots & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots \\ 0 & 0 & \cdots & r(T,1) & r(T-1,1) & \cdots & r(T-\tau_{\max},1) \end{bmatrix}, \qquad (3.10)$$

As before, this can be extended to the multivariate case of an *N*-channel system by replacing each column of **R** with *N* columns (each representing a separate recording channel). For *N* channels, the dimensions of **R** would be $T \times N\tau_{window}$. The constant term is included by concatenating *N* columns of ones to the left of **R**. In the context of speech, the stimulus variable, **s**, represents either a column-wise vector (e.g., envelope) or a $T \times F$ matrix (e.g., spectrogram). The resulting decoder, **g**, would be a vector of $N\tau_{window}$ samples or a $N\tau_{window} \times F$ matrix, respectively. While interpretation of decoder weights is not as straightforward as that of a TRF, one may wish to separate its dimensions (e.g., $N \times \tau_{window} \times F$) to examine the relative weighting of each channel. The channel weights represent the amount of information that each channel provides for reconstruction, i.e., highly informative channels receive weights of greater magnitude while channels providing little or no information receive weights closer to zero.

In Eq. 3.9, Tikhonov regularisation is used as it is assumed that the neural response data is multivariate. As mentioned in section 3.1, any bias from the correlation between the neural response channels is removed in the reconstruction approach. In practice, this is achieved by dividing out the autocovariance structure of the neural response (Eq. 3.9). As a result, channel weighting becomes much more localised because redundancies are no longer encoded in the model, giving it an advantage over the TRF method and cross-correlation approaches.

3.3 mTRF Toolbox: Implementation and Functionality

This section outlines how the reverse correlation method can be implemented in MATLAB using the mTRF Toolbox. Specifically, it describes how to train and test on univariate and multivariate datasets and how the resulting model should be optimised for specific purposes.

3.3.1 Training

Modelling the stimulus-response mapping of a given dataset is implemented in mTRF Toolbox using a simple function called *mTRFtrain* (Appendix B). This function computes univariate or multivariate ridge regression as described in the previous section (Eqs. 3.5, 3.6 and 3.9). The model can be trained on the data set in two separate ways: (1) by training on each trial separately and averaging over *M* models, or (2) by training on a concatenation of trials. Both of these approaches yield the same results because the data are modelled using a linear assumption. Here, the former approach will be considered because it affords certain advantages. Firstly, by generating separate models for each of the *M* trials, certain denoising algorithms that require repetition of "trials" can be applied to the model coefficients, even if they were modelled on different stimuli, e.g., joint decorrelation (de Cheveigné and Parra, 2014). Secondly, artefacts from discontinuities between trials are not an issue. Thirdly, cross-validation is much more efficient because training models on small amounts of data and averaging across trials is much faster than concatenating large amounts of data and training on them (see section 3.3.2).

For a given trial, the *mTRFtrain* function trains on all data features (e.g., frequency bands, response channels) simultaneously. The only requirement is that the stimulus and response data have the same sampling rate (which is specified in Hz) and be the same length in time. As described in the previous section, vectors and matrices should be organised such that all features are arranged column-wise. The mapping direction is specified as '1' (forward modelling) or '−1' (backward modelling). The minimum and maximum time lags are entered in milliseconds and converted to samples based on the sampling rate entered. It is often useful to include additional time lags such as prestimulus lags for visualisation purposes. The user should also account for regression artefacts at either extreme of the resulting model. However, when optimising models for prediction purposes, it is advisable to use only stimulus-relevant time lags. The lag matrix used in the ridge regression is generated by a function called *lagGen* (Appendix F). If the user specifies to map backwards, the lags are automatically reversed and the algorithm is changed from Eq. 3.5 to 3.9. If the stimulus entered is univariate (i.e., a vector), the algorithm will automatically switch to Eq. 3.6 to use the superior quadratic ridge penalty (see section 3.2.2). The final parameter that must be specified is the ridge parameter, λ. For visualisation of model coefficients, λ can be empirically chosen as the lowest value such that any increase would result in no visible improvement in the plotted estimate (Lalor et al., 2006). For optimising model performance, a more systematic approach should be implemented such as cross-validation, as described in the following section.

3.3.2 Optimisation

Optimisation of the stimulus-response mapping can be achieved via cross-validation and is implemented using the *mTRFcrossval* function (Appendix D). Specifically, the goal is to identify the value of the ridge parameter that optimises this mapping. Here, the entire dataset is entered together, with M stimuli and M response matrices arranged in two cell arrays. There is no requirement that the individual trials to be the same length in time (although this is preferable for optimisation reasons). Another important factor that optimises cross-validation is normalisation of both input and output data. By z-scoring the data, the range of values needed to conduct a comprehensive parameter search can be greatly reduced, making the processes more efficient. The ridge values over which validation is measured can be entered as a single vector. All other parameters are entered in the same way as in *mTRFtrain*.

The validation approach implemented in *mTRFcrossval* is that of ‘leave-one-out’ cross-validation, although this could also be described as M-fold cross-validation. First, a separate model is fit to each of the M trials for every ridge value specified. Then, the trials are rotated M times such that each trial is ‘left out’ or used as the ‘test set’, and the remaining $M-1$ trials are assigned as the ‘training set’ (see Fig. 3.2). The actual models tested are obtained by averaging over the single-trial models assigned to each training set. As mentioned earlier, this approach is more efficient than concatenating $M-1$ trials and fitting a model to these data. Each averaged model is then convolved with data from the corresponding test set to predict either the neural response (forward modelling) or the stimulus signal (backward modelling). This process is repeated for each of the ridge values. Validation of the model is assessed by comparing the predicted estimate with the corresponding original data. Two different validation metrics are used: Pearson’s correlation coefficient and mean squared error (MSE). Once the validation metrics have been obtained, they are averaged across all trials. This approach is advisable because each of the models should in theory require the same ridge value for regularisation, given that they share $M-2$ trials of data. This ensures that the models generalise well to new data and are not overfit to the test set. However, this approach works best if all the trials are the same length. The optimal ridge value is identified as that which yields either the highest *r*-value or the lowest MSE-score on average.

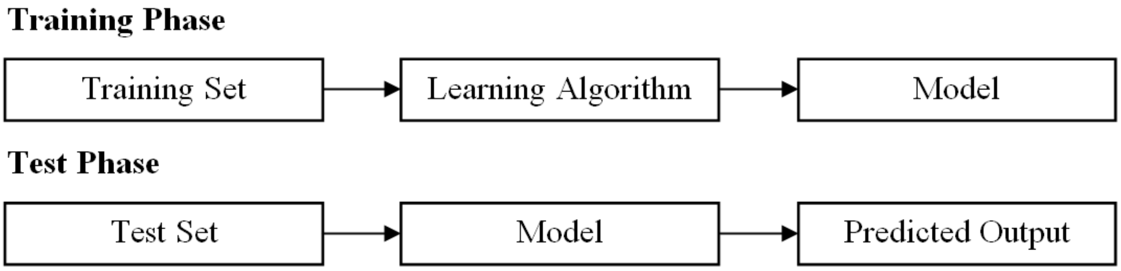


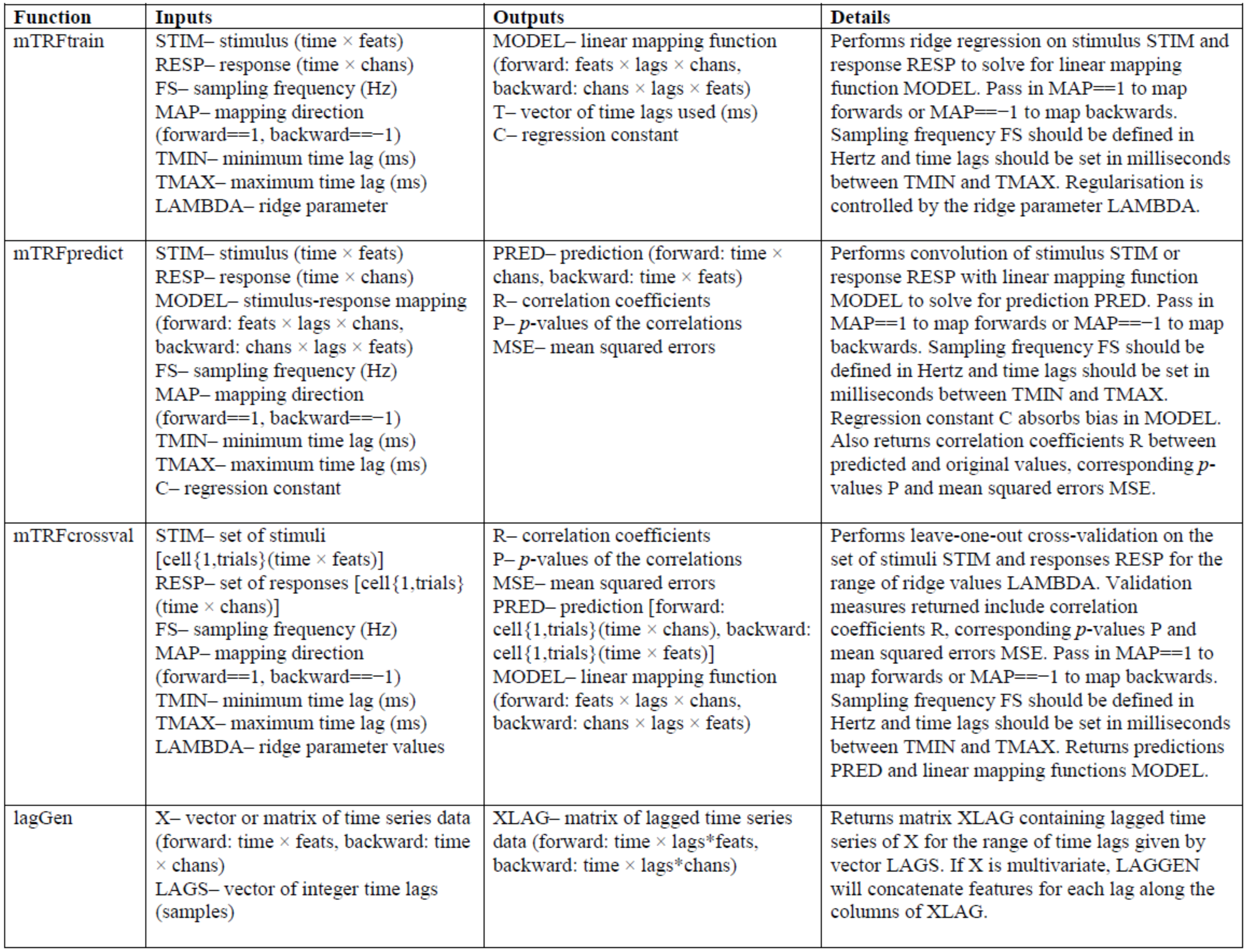
Figure 3.2: Optimisation procedure implemented by *mTRFcrossval*.

3.3.3 Testing

Once the model parameters have been tuned using cross-validation, the optimised model can be tested on new data using the *mTRFpredict* function (Appendix C). This can be conducted on data that was held aside from the cross-validation procedure (which is considered good practice) or on the same test data used for cross-validation. As previously mentioned, because the above cross-validation procedure takes the average of the validation metric across trials, the models are not biased towards the test data used for cross-validation. Thus, it is legitimate to report model performance based on these data because testing on new unseen data will likely yield the same result.

While the *mTRFpredict* function outputs the same performance metrics as *mTRFcrossval*, it also outputs the predicted signal for further evaluation. When predicting a multivariate signal such as EEG, a performance measure is calculated for every feature (i.e., EEG channel), allowing the user to base evaluation of the model on whichever features they deem most relevant. For a summary of the functions in mTRF Toolbox, please refer to Table 3.1.

Table 3.1: Summary of mTRF Toolbox functions.



Function	Inputs	Outputs	Details
mTRFtrain	STIM– stimulus (time × feats) RESP– response (time × chans) FS– sampling frequency (Hz) MAP– mapping direction (forward==1, backward==−1) TMIN– minimum time lag (ms) TMAX– maximum time lag (ms) LAMBDA– ridge parameter	MODEL– linear mapping function (forward: feats × lags × chans, backward: chans × lags × feats) T– vector of time lags used (ms) C– regression constant	Performs ridge regression on stimulus STIM and response RESP to solve for linear mapping function MODEL. Pass in MAP==1 to map forwards or MAP==−1 to map backwards. Sampling frequency FS should be defined in Hertz and time lags should be set in milliseconds between TMIN and TMAX. Regularisation is controlled by the ridge parameter LAMBDA.
mTRFpredict	STIM– stimulus (time × feats) RESP– response (time × chans) MODEL– stimulus-response mapping (forward: feats × lags × chans, backward: chans × lags × feats) FS– sampling frequency (Hz) MAP– mapping direction (forward==1, backward==−1) TMIN– minimum time lag (ms) TMAX– maximum time lag (ms) C– regression constant	PRED– prediction (forward: time × chans, backward: time × feats) R– correlation coefficients P– *p*-values of the correlations MSE– mean squared errors	Performs convolution of stimulus STIM or response RESP with linear mapping function MODEL to solve for prediction PRED. Pass in MAP==1 to map forwards or MAP==−1 to map backwards. Sampling frequency FS should be defined in Hertz and time lags should be set in milliseconds between TMIN and TMAX. Regression constant C absorbs bias in MODEL. Also returns correlation coefficients R between predicted and original values, corresponding *p*-values P and mean squared errors MSE.
mTRFcrossval	STIM– set of stimuli [cell{1,trials}(time × feats)] RESP– set of responses [cell{1,trials}(time × chans)] FS– sampling frequency (Hz) MAP– mapping direction (forward==1, backward==−1) TMIN– minimum time lag (ms) TMAX– maximum time lag (ms) LAMBDA– ridge parameter values	R– correlation coefficients P– *p*-values of the correlations MSE– mean squared errors PRED– prediction [forward: cell{1,trials}(time × chans), backward: cell{1,trials}(time × feats)] MODEL– linear mapping function (forward: feats × lags × chans, backward: chans × lags × feats)	Performs leave-one-out cross-validation on the set of stimuli STIM and responses RESP for the range of ridge values LAMBDA. Validation measures returned include correlation coefficients R, corresponding *p*-values P and mean squared errors MSE. Pass in MAP==1 to map forwards or MAP==−1 to map backwards. Sampling frequency FS should be defined in Hertz and time lags should be set in milliseconds between TMIN and TMAX. Returns predictions PRED and linear mapping functions MODEL.
lagGen	X– vector or matrix of time series data (forward: time × feats, backward: time × chans) LAGS– vector of integer time lags (samples)	XLAG– matrix of lagged time series data (forward: time × lags*feats, backward: time × lags*chans)	Returns matrix XLAG containing lagged time series of X for the range of time lags given by vector LAGS. If X is multivariate, LAGGEN will concatenate features for each lag along the columns of XLAG.

3.4 Examples

The examples presented in this section use data from a published study that measured EEG responses of human participants to natural, continuous speech (Di Liberto et al., 2015). The subject listened to an audiobook version of a classic work of fiction read by a male American English speaker. The audio was presented in 28 segments (each ~155 s in duration), of which a subset of five are used in the examples in this chapter. EEG data were recorded using a 128-channel ActiveTwo system (BioSemi) and digitised at a rate of 512 Hz. Offline, the data were digitally filtered between 1–15 Hz, downsampled to a rate of 128 Hz and re-referenced to the left and right mastoid channels. Only 32 of the 128 channels recorded are included in the analysis, but crucially, are distributed evenly across the head (Mirkovic et al., 2015). Further details can be found in the original study (Di Liberto et al., 2015).

This section details several examples that demonstrate how the mTRF Toolbox can be utilised to relate neural data to sensory stimuli in a variety of different ways. These include:

1. Univariate TRF estimation
2. Optimisation and prediction
3. Multivariate TRF analysis
4. Stimulus reconstruction
5. TRF versus cross-correlation

3.4.1 Univariate TRF Estimation

The aim here is to estimate the temporal response function that maps a univariate representation of the speech envelope onto the EEG signal recorded at each channel. The broadband envelope of the speech signal (Fig. 3.3*A*) was calculated using:

$$x_a(t) = x(t) + j\hat{x}(t), \tag{3.11}$$

where $x_a(t)$ is the complex analytic signal obtained by the sum of the original speech $x(t)$ and its Hilbert transform $\hat{x}(t)$. The envelope was defined as the absolute value of $x_a(t)$. This was then downsampled to the same sampling rate as the EEG data, after applying a zero-phase shift anti-aliasing filter. TRFs were calculated between lags of −150 and 450 ms, allowing an additional 50 ms at either end for regression artefacts. An estimate was computed separately for each of the five trials and then averaged. The ridge parameter was empirically chosen to maintain component amplitude (Lalor et al., 2006).

A measure of global field power (GFP) was first estimated by calculating TRF variance across the 32 channels (Fig. 3.3*B*). GFP constitutes a reference-independent measure of response strength across the entire scalp at each time lag (Lehmann and Skrandies, 1980, Murray et al., 2008). Based on the temporal profile of the GFP measure, two dominant TRF components were identified at ~80 and ~140 ms. Fig. 3.3*C* shows the scalp topographies of each of these components. Their latency and polarity resemble that of the classic N1 and P2 components of a typical (mastoid-referenced) auditory-evoked response (Stekelenburg and Vroomen, 2007). The topography of the N1-P2 complex suggests that both components are strongest at fronto-central position FCz. The grand average TRF calculated at FCz is shown in Fig. 3.3*D*, along with the TRF measured at occipital location Oz for comparison.

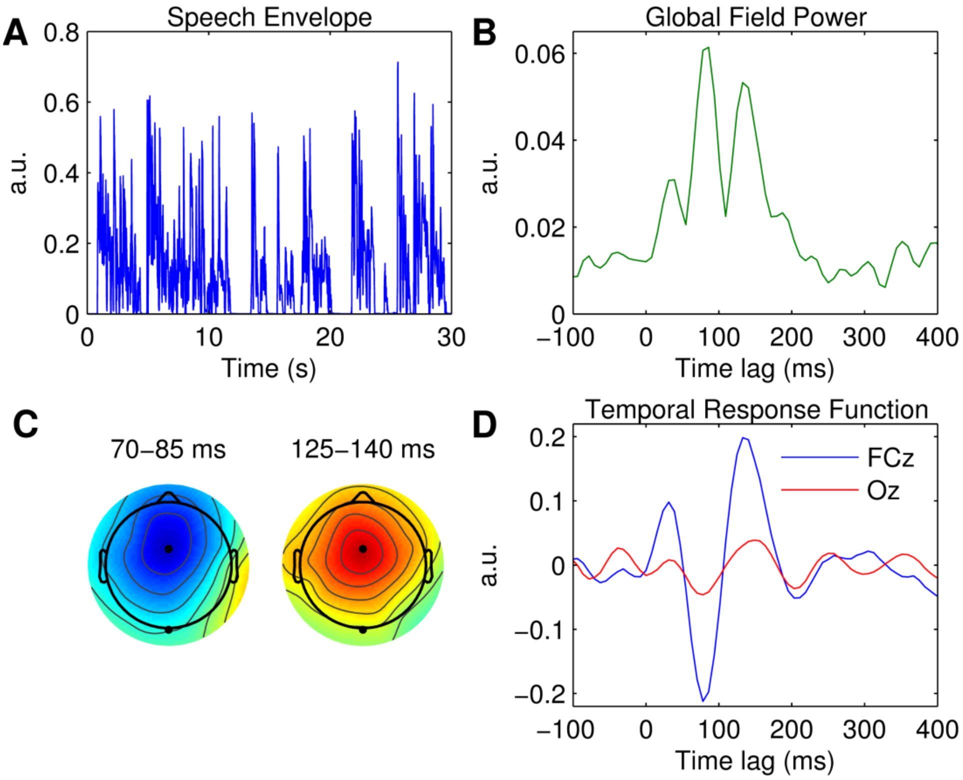


Figure 3.3: Univariate TRF estimation.

A, A 30-second segment of the broadband speech envelope. ***B***, Global field power measured at each time lag. ***C***, Scalp topographies of the dominant TRF components occurring at 78 ms and 141 ms. The black markers indicate the locations of fronto-central channel, FCz, and occipital channel, Oz. ***D***, Grand average TRFs at FCz (blue trace) and Oz (red trace).

3.4.2 Optimisation and Prediction

The aim here is to use the TRF model to predict the EEG response of unseen data. This time, tuning of model parameters was conducted using a more systematic approach, i.e., that of the cross-validation procedure described earlier (see section 3.3.3). Specifically, TRFs were calculated for a range of ridge values $\left(\lambda = 2^0, 2^2, ..., 2^{20}\right)$ on each of the separate trials. For each ridge value, the TRFs were averaged across every combination of four trials and used to predict the EEG of the remaining fifth trial. Here, the data were modelled at time lags between 0–200 ms as these lags reflected the most information in the global TRF responses (Fig. 3.3*B*). Inclusion of additional lags (pre-stimulus or post-stimulus) did not improve the model estimate.

Fig. 3.4*A* shows the results of the cross-validation based on the correlation coefficient (Pearson's *r*) between the original and predicted EEG responses. Critically, the *r*-values were averaged across the five trials to prevent overfitting the model to the test data. The *r*-values were also averaged across the 32 channels such that model performance would be optimised in a more global manner. Alternatively, one could average across only channels within a specified top percentile or based on a specific location. Fig. 3.4*B* shows the results of the cross-validation based on the mean squared error (MSE). The same averaging procedure was used to identify the optimal ridge value here.

The ridge value was chosen such that it maximised the correlation between the original and predicted EEG (David and Gallant, 2005). Note that using MSE as a criteria for cross-validation would have yielded the same result. Fig. 3.4*C* shows the correlation coefficient obtained at each channel using the optimised TRF model. The topographical distribution of Pearson's *r* is very similar to that of the dominant TRF components (Fig. 3.3*C*). Indeed, it is unsurprising that the model performed best at channels where the response was strongest. Fig. 3.4*D* shows two-second segments of the EEG response at FCz and the corresponding estimate predicted by the optimised TRF model.

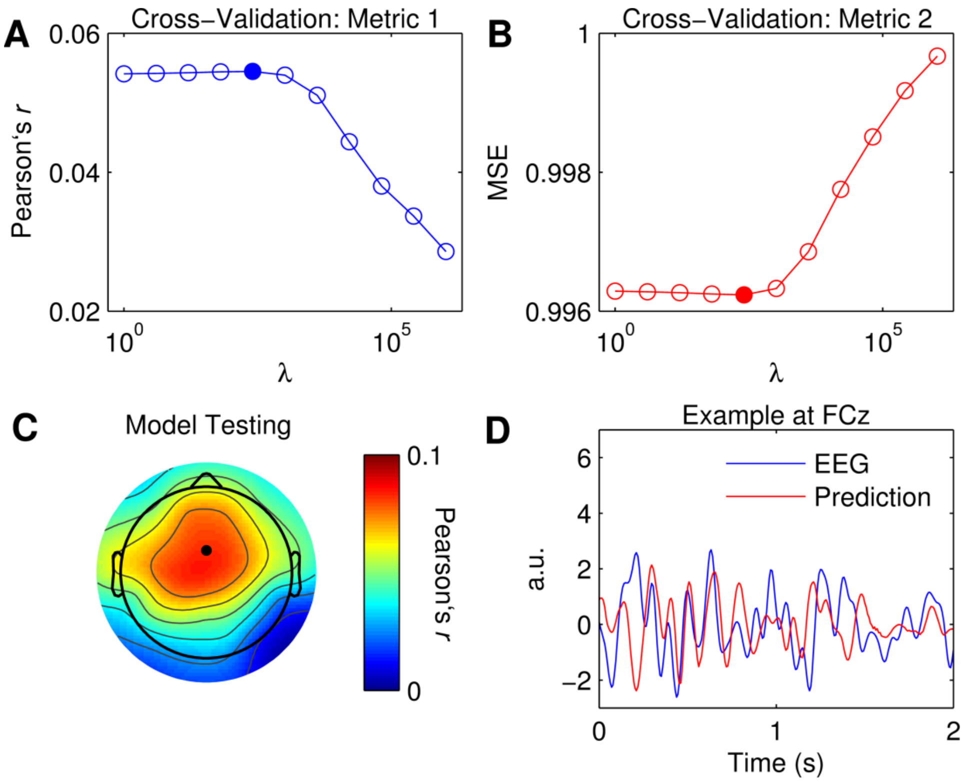


Figure 3.4: Optimisation of TRFs for EEG prediction.

A, Cross-validation of model based on the correlation between the original and predicted EEG response (Pearson's *r* averaged across channels and trials). The filled marker indicates the highest *r*-value, i.e., the optimal ridge value. ***B***, Cross-validation based on mean squared error (MSE). The optimal ridge value is identified by the lowest MSE-score. ***C***, Test of the optimised TRF model shows the correlation coefficient at each channel. The black marker indicates the location of channel FCz. ***D***, Two-second segments of the EEG response at FCz (blue trace) and the corresponding estimate predicted by the optimised TRF model (red trace).

3.4.3 Multivariate TRF Analysis

The aim here is to estimate the TRF for a multivariate (spectrogram) representation of speech, i.e., an mTRF. The spectrogram representation (Fig. 3.5*A*) was obtained by first filtering the speech stimulus into 16 logarithmically-spaced frequency bands between 250 Hz and 8 kHz according to Greenwoods equation (Greenwood, 1990). Filtering the data in a logarithmic manner attempts to model the frequency analysis performed by the auditory periphery (see section 2.1.2). The energy in each frequency band was calculated using a Hilbert transform as above (Eq. 3.11).

For visualisation, mTRFs were calculated between lags of −150 and 450 ms and model parameters were tuned empirically. Fig. 3.5*B* shows the mTRF response at channel FCz for all frequency bands between 0.25–8 kHz. Visual inspection of Fig. 3.5*B* suggests that the dominant $\mathrm{N1}_{\mathrm{TRF}}$ and $\mathrm{P2}_{\mathrm{TRF}}$ components encoded speech information at nearly every frequency band up to ~6 kHz, which is where most of the information was contained in the speech signal (see Fig. 3.5*A*). Averaging the mTRF across frequency bands would yield a univariate TRF measure that closely approximates the TRF calculated using the broadband envelope (Fig. 3.3*D*).

To predict the EEG response with the mTRF model, the same approach was implemented as before. Although the results yielded by the cross-validation (Fig. 3.5*C*,*D*) were similar to those for the univariate TRF approach (Fig. 3.4*A*,*B*), the mTRF approach appeared to be more sensitive to changes in the ridge value. Further investigation revealed that this could not be attributed to using different regularisation penalties in univariate and multivariate analyses (see section 3.2.2). Despite this, performance of the optimised mTRF model was akin to that of the univariate TRF model over the entire scalp (Fig. 3.5*E*,*F*).

While it has been demonstrated that mTRF models are superior to univariate TRF models for predicting EEG responses (Di Liberto et al., 2015), it must be taken into consideration that multivariate TRF analysis is more sensitive to regularisation (certainly for ridge regression) and can involve considerably more computations.

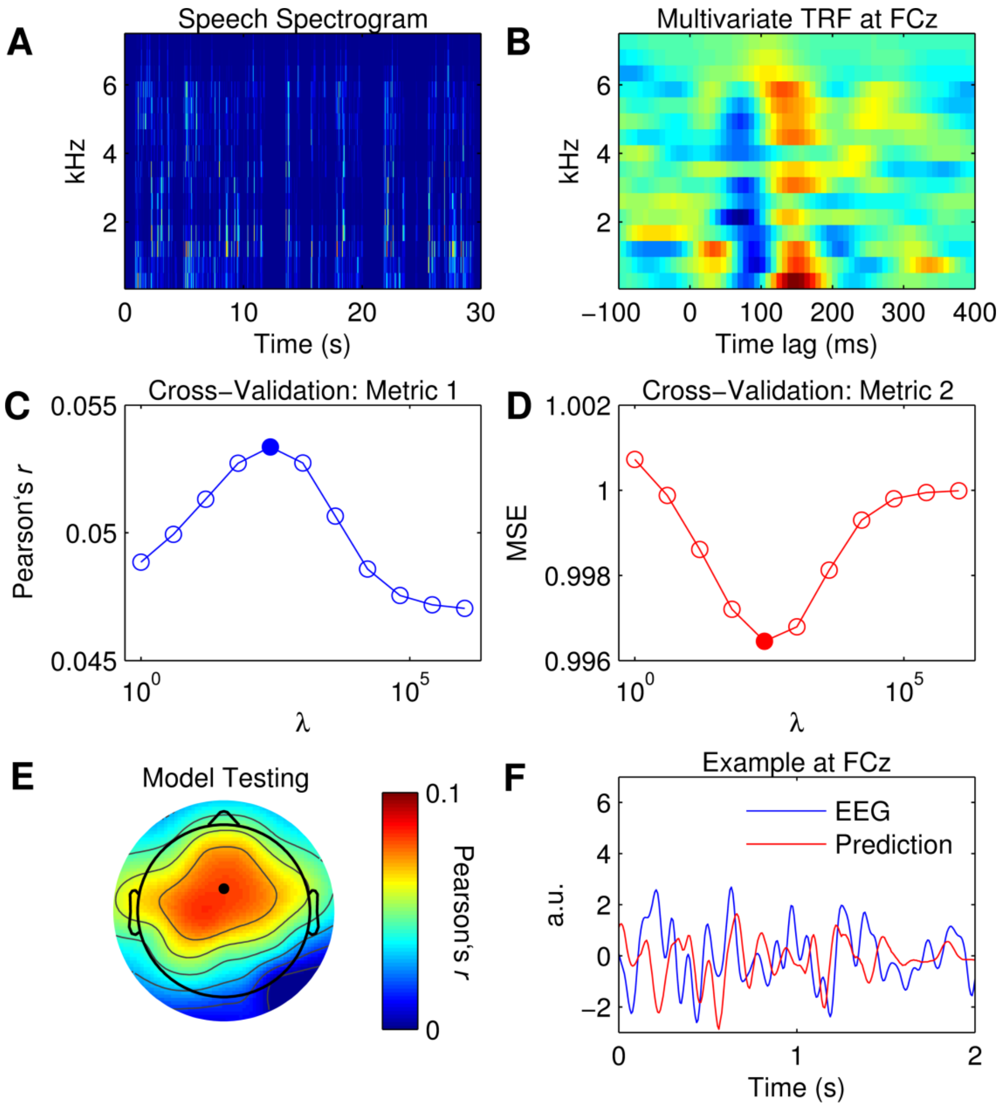


Figure 3.5: Multivariate TRF estimation and EEG prediction.

A, A 30-second segment of the speech spectrogram. ***B***, Grand average mTRF at channel FCz. ***C***, Cross-validation of model based on the correlation between the original and predicted EEG response (Pearson's *r* averaged across channels and trials). The filled marker indicates the highest *r*-value, i.e., the optimal ridge value. ***D***, Cross-validation based on mean squared error (MSE). The optimal ridge value is identified by the lowest MSE-score. ***E***, Test of the optimised mTRF model shows the correlation coefficient at each channel. The black marker indicates the location of channel FCz. ***F***, Two-second segments of the EEG response at FCz (blue trace) and the corresponding estimate predicted by the optimised TRF model (red trace).

3.4.4 Stimulus Reconstruction

The aim here is to generate a decoder that models the data in the backwards direction (i.e., from EEG to stimulus) and to use it to reconstruct an estimate of the univariate stimulus input. The advantages of this approach over the forward modelling technique are outlined in section 3.1. Tuning of model parameters was conducted using the same cross-validation technique described for the TRF models (see section 3.4.2). Specifically, decoders were calculated for the same range of ridge values $\left(\lambda = 2^0, 2^2, ..., 2^{20}\right)$ at time lags between 0–200 ms. The difference here was that the EEG was treated as the "input" and the stimulus as the "output", and the direction of the lags was reversed, i.e., −200 to 0.

Fig. 3.6*A* shows the results of the cross-validation as measured by the correlation coefficient between the original and reconstructed speech envelope, while Fig. 3.6*B* represents validation of the model ridge parameter based on MSE. Again, both metrics have been averaged across trials to prevent overfitting to the test data. All 32 EEG channels were included in the model validation procedure to optimise performance (see section 3.1). The advantages of the backward modelling approach over forward modelling are evidenced by the dramatic reduction in residual error as indexed by both the *r*-values and MSE-scores. This is mainly attributable to the fact that the decoder can utilise information across the entire head simultaneously (i.e., in a multivariate sense) to determine the speech estimate, whereas when modelling in the forward direction, the predicted EEG estimate is based on a single univariate mapping between the stimulus and the EEG response at that specific channel (Mesgarani et al., 2009).

While the decoder channel weights are not readily interpretable in a neurophysiological sense, their weighting reflects the channels that contribute most towards reconstructing the stimulus signal (Haufe et al., 2014). Fig. 3.6*C* shows the decoder weights averaged across time lags between 110–130 ms (this was close to where weighting was maximal as indexed by GFP). In comparison to the TRF topographies (Fig. 3.3*C*), the distribution of model weight is much more localised and, interestingly, right lateralised. Because the decoder is not required to encode information at every channel across the scalp as a TRF does, it can selectively weight only those channels important for reconstruction, whilst ignoring irrelevant channels by giving them zero weight (Haufe et al., 2014). A two-second sample of a reconstructed estimate can be seen in Fig. 3.6*D*. Stimulus reconstruction for a multivariate stimulus is conducted in much the same manner, except model performance must be evaluated for every feature (e.g., frequency band) separately or by averaging across features and then evaluating.

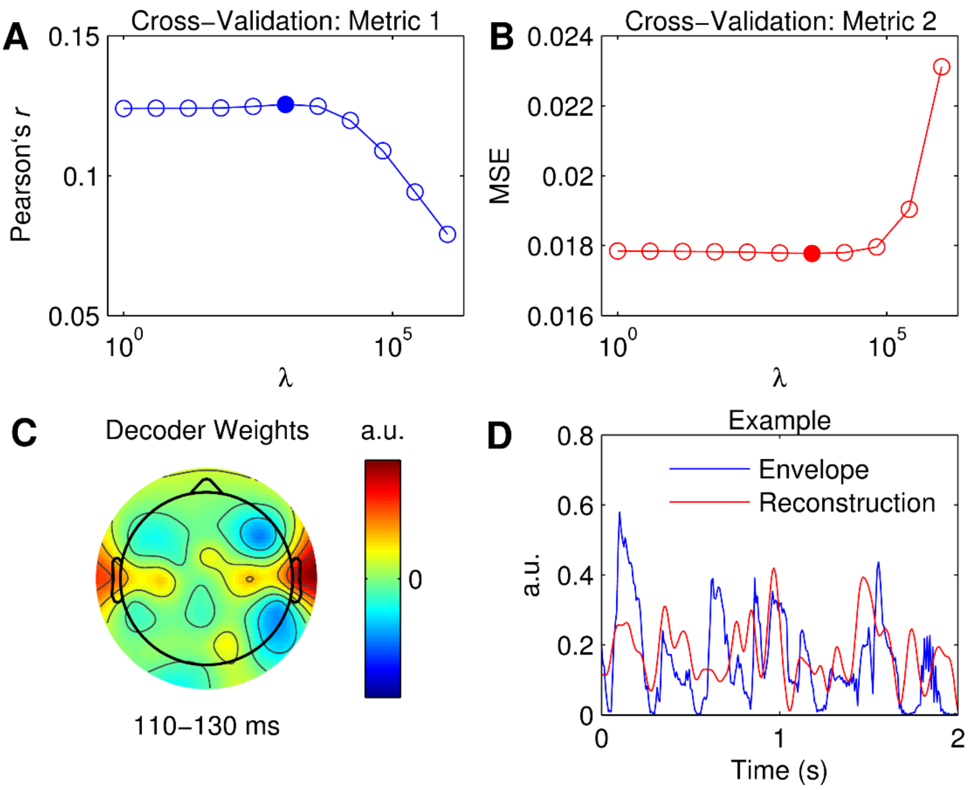


Figure 3.6: Stimulus reconstruction.

A, Cross-validation of model based on the correlation between the original and reconstructed speech envelope (Pearson's *r* averaged across trials). The filled marker indicates the highest *r*-value, i.e., the optimal ridge value. ***B***, Cross-validation based on mean squared error (MSE). The optimal ridge value is identified by the lowest MSE-score. ***C***, Decoder channel weights averaged over time lags between 110–130 ms. ***D***, Two-second segments of the original speech envelope (blue trace) and the corresponding estimate reconstructed by the optimised decoder (red trace).

3.4.5 TRF versus Cross-Correlation

As mentioned earlier, the impulse response of an LTI system can be easily approximated via a simple cross-correlation of the input and output. While this approach is much more straightforward than using ridge regression, it is only suitable for input signals that conform to a stochastic process. To demonstrate this empirically, a comparison is made between each of these approaches using both speech and white noise as a stimulus input signal. The speech data presented here are the same as those in the previous examples. The non-speech data presented here were published in a study that investigated the TRF approach for estimating the response of the auditory system to Gaussian white noise (Lalor et al., 2009). The subject listened to ten 120-s segments of uninterrupted noise stimuli, of which a subset of six are used in this example. The stimuli were Gaussian broadband noise with energy limited to a bandwidth of 0–22.05 kHz, modulated using Gaussian noise signals with uniform power in the range 0–30 Hz. To account for the logarithmic nature of auditory stimulus intensity perception, the values of these modulating signals, x, were then mapped to the amplitude of the audio stimulus, x', using the following exponential relationship:

$$x' = 10^{2x}. \tag{3.12}$$

EEG data were recorded and processed using the exact same procedure described in the previous examples (see section 3.4). Further details can be found in the original study (Lalor et al., 2009).

Examples of the speech and noise stimuli used in the experiments are shown in Fig. 3.7*A* and Fig. 3.7*B* respectively. The autocorrelation of each stimulus reveals that the speech stimulus is correlated with itself at different lags (Fig. 3.7*C*), whereas the noise stimulus is only with itself at zero time lag (Fig. 3.7*D*). Fig. 3.7*F* shows the impulse response for the white noise stimulus calculated at channel FCz using the TRF approach and a cross-correlation approach. Visual inspection suggests that the cross-correlation approximated the impulse response as accurately as the TRF approach. However, the same was not true for the speech stimulus, where the cross-correlation visibly smeared the impulse response across time compared to the TRF approach (Fig. 3.7*E*). This demonstrates the utility of the mTRF method for characterisation of sensory systems in response to naturalistic stimuli such as speech.

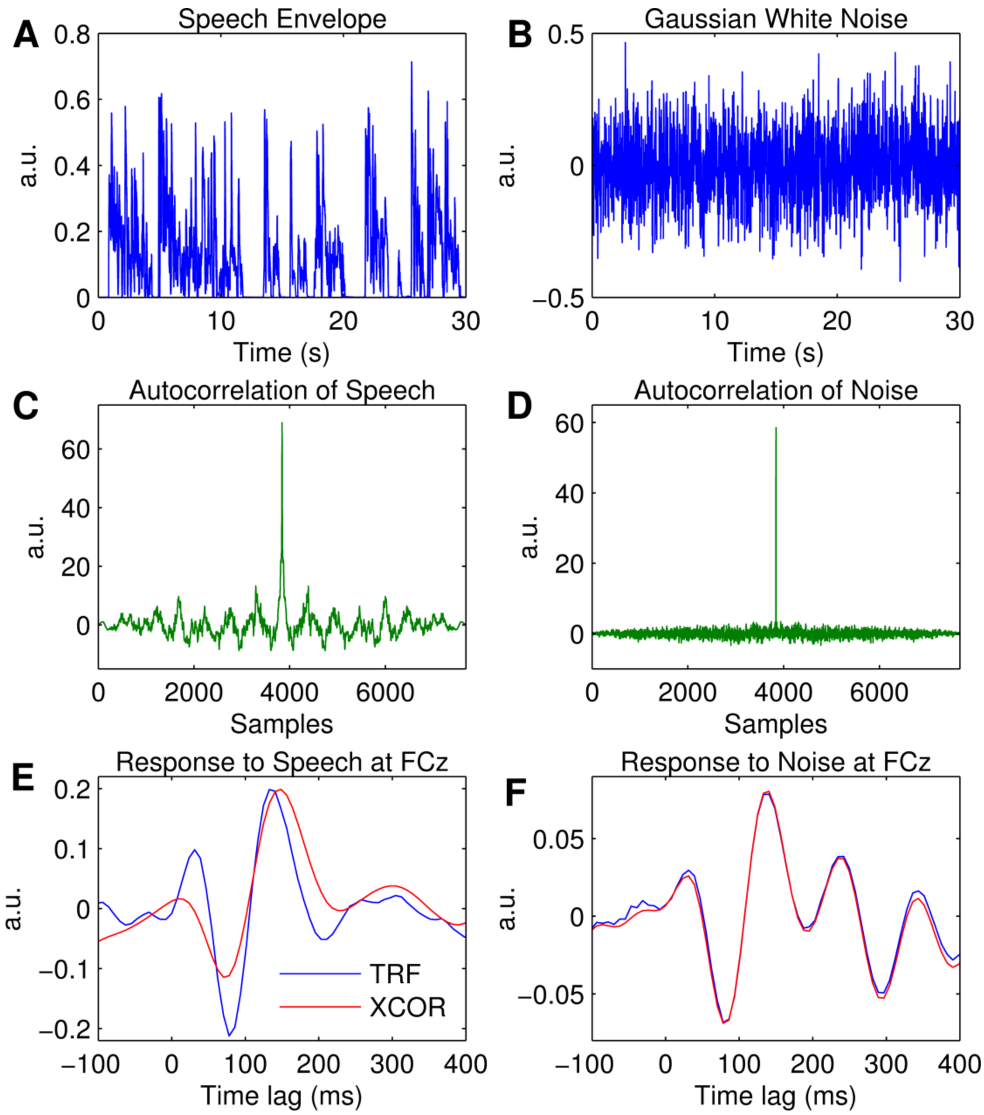


Figure 3.7: Comparison of the TRF and cross-correlation (XCOR) approach.

A, A 30-second segment of the broadband speech envelope. ***B***, A 30-second segment of amplitude modulated noise. ***C***, Autocorrelation of the speech envelope. ***D***, Autocorrelation of the noise signal. ***E***, The impulse response to speech at channel FCz estimated using the TRF approach (blue trace) and the cross-correlation approach (red trace). ***F***, The impulse response to white noise at channel FCz estimated using the TRF approach (blue trace) and the cross-correlation approach (red trace).

3.5 Discussion

This chapter has described a new MATLAB-based toolbox for modelling the relationship between neural signals and natural, continuous stimuli. The previous examples demonstrate how this versatile toolbox can be applied to both univariate and multivariate datasets, to map in both the forwards and backwards direction. The advantages of using this approach over traditional ERP and cross-correlation methods have also been demonstrated.

3.5.1 Applications

The mTRF Toolbox has many applications in sensory neuroscience, none more so than for studying how natural, continuous speech is processed in the human brain. The forward TRF method has already been successfully applied in several human speech studies (Lalor and Foxe, 2010, Power et al., 2012, Di Liberto et al., 2015). This work has yielded several key findings relating to how the brain selectively attends to a single speech stream in a cocktail party scenario (Power et al., 2012) and how spectrotemporal and phonetic information are represented in auditory cortical activity (Di Liberto et al., 2015). Several other groups have implemented similar forward and backward modelling approaches to study auditory scene analysis in humans (Ding and Simon, 2012a, Mesgarani and Chang, 2012, Zion-Golumbic et al., 2013b). However, there has been less focus on studying multisensory speech using an SI approach. One of the main reasons for this is the difficulty involved in disentangling contributions from multisensory interactions with those of unisensory processing. One study which examined AV speech using a forward TRF approach showed that an early MEG component at ~50 ms was enhanced by AV speech compared to audio-only speech. However, they did not account for unisensory visual contributions in auditory cortex, thus the effect may not entirely reflect multisensory processing. There is a need to develop a reliable framework for applying SI techniques to multisensory data, particularly in the context of a stimulus reconstruction approach.

Aside from studying speech, the forward TRF approach has been extensively applied in vision research to study how the brain processes stimuli that modulate in contrast over time (Lalor et al., 2006, Murphy et al., 2012, Frey et al., 2010, Lalor et al., 2007). This particular approach has also been applied in clinical research to investigate visual processing deficits in children with ASD (Frey et al., 2013) and in adults with schizophrenia (Lalor et al., 2012, Lalor et al., 2008). More recently, it has been extended to studying how the brain processes more natural visual stimuli such as coherent motion (Gonçalves et al., 2014). Despite the versatility of this approach and its obvious utility in clinical research, there are many aspects of human vision research that have not yet been explored in this way, and in particular, using the stimulus reconstruction method. Although stimulus reconstruction has not been widely used in human vision research, it has been successfully used to decode finger movements from surface EMG signals (Krasoulis et al., 2015), thus further demonstrating its versatility.

3.5.2 Considerations

The linear assumption underlying the reverse correlation method has implications for its interpretation. This assumption of a linear relationship between stimulus intensity and neural response amplitude likely results in a response measure reflective of feedforward activity in a subset of cortical cells (Lalor et al., 2009). Thus, it is possible that such an approach is insensitive to cortical responses that relate to the stimulus in a non-linear manner including lateral and feedback contributions, which may have implications for studying multisensory integration. This is in contrast to the challenge involved in disambiguating the myriad feedforward, lateral and feedback contributions to the time-locked average ERP (Di Russo et al., 2005).

These linear assumptions will need to be addressed in future work in order to accurately characterise populations of neurons that respond in a non-linear way to complex stimuli (Theunissen et al., 2000). Previous work has already developed a quadratic extension of the linear TRF approach for modelling visual responses to contrast stimuli, but did not find any significant improvement in model performance relative to that of a linear model (Lalor et al., 2008). However, subsequent studies that applied the same quadratic model to the auditory system demonstrated marginal improvements in model performance for acoustic white noise stimuli (Power et al., 2011a, Power et al., 2011b). Expansion of the TRF model into higher orders has also been explored using machine learning techniques such as support vector regression (Crosse, 2011). However, while such an approach can lead to marginal improvements in model performance, the increased computational complexity makes it very unsuitable for use on large data sets such as EEG.

The fact that non-linear models perform only marginally better than linear models for population data (e.g., EEG; Power et al., 2011a, Power et al., 2011b), and yet appear to be more beneficial for modelling single-unit data (e.g., ECoG; Theunissen et al., 2000) may imply something fundamental about the nature of EEG data. Each EEG electrode measures cortical activity from a large neural population due to the effects of volume conduction (and to a lesser extent the electrode surface area; see section 2.2).

Thus, it can only detect neural activity that is encoded by the entire (or most of the) population it is recording from; everything else becomes “noise” and cancels each other out via superposition. It is possible that linear activity is encoded more globally by neural populations and that non-linear activity is more localised to smaller sub-populations of neurons. This theory is supported by the diversity of non-linear responses across neurons (Mesgarani et al., 2009). The effect of volume condition would result in non-linear activity being cancelled out and thus being undetected by the EEG recording. If this were the case, then using a linear model to approximate the response of a neural population would be all the more justifiable.

While the brain certainly does not possess the properties of an LTI system, there are distinct advantages to treating it as one in certain circumstances, as evidenced by the above examples and the vast neurophysiology literature. The Toolbox described in this chapter provides a straightforward way to model any sensory system as if it were LTI and is implemented in several AV speech studies in the following chapters of this thesis.

Chapter 4 Congruent Visual Speech Enhances Cortical Entrainment to Continuous Auditory Speech in Noise-Free Conditions

4.1 Introduction

During natural, everyday conversation, we routinely process speech using both our auditory and visual systems. The benefit of viewing a speaker's articulatory movements for speech comprehension has been well documented and has been characterised in terms of two specific modes of audiovisual information: 'complementary' and 'correlated' (see section 2.3.2; Summerfield, 1987, Campbell, 2008, Grant and Seitz, 2000). Visual speech assumes a complementary role when it is required to compensate for underspecified auditory speech, enhancing perception, for example, in adverse hearing conditions (Ross et al., 2007a, Sumby and Pollack, 1954) and in people with impaired hearing (Grant et al., 1998). It assumes a correlated role when there is redundancy between the information provided by vision and audition, for example, in optimal listening conditions where it has been shown to benefit people with normal hearing (Reisberg et al., 1987). Specifically, in the latter case, enhanced perception is possible because the visible articulators that determine the vocal resonances, such as the lips, teeth and tongue, as well as ancillary movements, such as facial, head and hand movements, are temporally correlated with the vocalised acoustic signal (see section 2.3.1; Summerfield, 1992, Jiang and Bernstein, 2011, Grant and Seitz, 2000). However, relatively little research has explicitly examined how the temporal correlation between auditory and visual speech impacts upon the neural processing of continuous AV speech.

EEG and MEG studies have demonstrated that auditory cortical activity entrains to the temporal envelope of speech (Abrams et al., 2008, Ahissar et al., 2001, Aiken and Picton, 2008, Lalor and Foxe, 2010). While many studies have examined the effects of attention on envelope tracking (Ding and Simon, 2012a, Power et al., 2012, O'Sullivan et al., 2015), less work has examined how this process may be influenced by visual speech (but see Zion-Golumbic et al., 2013a, Luo et al., 2010). Traditionally, EEG/MEG studies have focused on how the brain responds to discrete AV stimuli such as syllables (Sams et al., 1991, Möttönen et al., 2004, Besle et al., 2004a, Möttönen et al., 2002), an approach that is limited in what it can say about the role of the temporal correlation between continuous auditory and visual speech. Indeed, many EEG/MEG studies have reported interesting crossmodal interaction effects on cortical response measures, even when the discrete stimuli were phonetically incongruent (Klucharev et al., 2003, van Wassenhove et al., 2005, Stekelenburg and Vroomen, 2007, Arnal et al., 2009). This is unsurprising, given that particular combinations of incongruent AV syllables elicit illusory percepts when presented concurrently (McGurk and MacDonald, 1976). It has been suggested (Campbell, 2008) that because such discrete incongruent stimuli are spatially and temporally coherent and coextensive, this may act as a cue to their integration (see section 2.4.2).

In this chapter, natural, continuous speech stimuli were used to examine how EEG entrains to temporally and contextually congruent and incongruent AV speech. Specifically, it was hypothesised that the benefits of congruent AV speech will be detectable in noise-free conditions and indexed by enhanced envelope tracking. Several follow-up experiments were implemented to answer the following research questions: (1) Is a dynamic human face sufficient to enhance envelope tracking, even when it is temporally incongruent? (2) Does contextually incongruent information, such as conflicting gender, modulate envelope tracking differently? (3) Is any dynamic visual stimulus sufficient to enhance envelope tracking, even if it does not comprise a human face? (4) Conversely, does a static human face enhance the tracking of a dynamic auditory input?

To obtain a direct measure of envelope tracking, the stimulus reconstruction approach (Mesgarani et al., 2009) was implemented using the mTRF Toolbox. One of the main goals of this chapter was to establish a framework for quantifying multisensory interactions using stimulus reconstruction. Within this framework, modulations in multisensory integration across different timescales were examined with a view to elucidating whether the effects were more prominent at any particular level of speech processing (i.e., phonemic, syllabic, word, prosodic; Giraud and Poeppel, 2012). The results of this study were presented at the *15th International Multisensory Research Forum* in Amsterdam in June, 2014 (Appendix G) and published in *The Journal of Neuroscience* (Crosse et al., 2015a).

4.2 Methods

4.2.1 Participants

Twenty-one native English speakers (8 females; age range: 19–37 years) participated in the experiment. Written informed consent was obtained from each participant beforehand. All participants were right-handed, free of neurological diseases, had self-reported normal hearing and normal or corrected-to-normal vision. The experiment was undertaken in accordance with the Declaration of Helsinki and was approved by the Ethics Committee of the Health Sciences Faculty at Trinity College Dublin, Ireland.

4.2.2 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, centred in the frame (see Fig. 4.1). Speech was directed at the camera and the speaker used frequent, but natural, hand movements. There was no background movement or noise. The speech was conversational-like and continuous, with no prolonged pauses between sentences. The linguistic content centred on political policy and the language was colloquial American English. Fifteen 60-s videos were rendered into 1280 × 720-pixel movies in VideoPad Video Editor (NCH Software). Each video had a frame rate of 30 frames/s and the soundtracks were sampled at 48 kHz with 16-bit resolution. Dynamic range compression was applied to each soundtrack in Audacity audio editor such that lower intensities of the speech signal could be amplified. Compression was applied at a ratio of 10:1 above a threshold of −60 dB. The signal was only amplified above a noise floor of −45 dB which prevented the gain increasing during pauses and unduly amplifying breathing sounds. The intensity of each soundtrack, measured by root mean square (RMS), was normalised in MATLAB (MathWorks).

To test the main hypothesis of the study and the four follow-up questions posed in the introduction, the same 15 soundtracks were dubbed to five different kinds of visual stimulus: (1) Congruent audiovisual stimuli (AVc) were created by re-dubbing each soundtrack to its original video, i.e., A1V1, A2V2, etc. Unimodal versions were also produced as a control, i.e., audio-only stimuli (A) and visual-only stimuli (V). (2) To examine the role of temporal congruency, incongruent audiovisual stimuli (AVi) were created by mismatching the same 15 soundtracks and videos, i.e., A1V2, A2V3, etc. (3) To examine the role of contextual congruency, the soundtracks were dubbed to videos of incongruent female speakers (AVif). The female speakers were centred in the frame (head, shoulders and chest) and their speech was directed at the camera. (4) To examine the impact of a dynamic (non-human) visual stimulus, incongruent nature stimuli (AVin) were created by dubbing the speech soundtracks to wildlife documentaries. (5) To examine the role of human-specific visual features, the soundtracks were dubbed to still images of the male speaker's static face (AVsf). For a summary of all the stimuli used in the experiment, please refer to Table 4.1.

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. 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). Each of the 15 speech passages was presented seven times, each time as part of a different experimental condition (see Table 4.1). Presentation order was randomised across conditions, within participants. Participants were instructed to fixate on either the speaker's mouth (V, AVc, AVi, AVif, AVsf) or on a grey crosshair (A, AVin) and to minimise eye blinking and all other motor activity during recording.

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. All target words were detectable in the auditory modality except during the V condition, where they were only visually detectable. Hits were counted for responses that were made 200–2000 ms after the onset of auditory voicing and feedback was given at the end of each trial. A target word could occur between one and three times in a given 60-s trial and there were exactly 30 targets in total per condition. A different set of target words was used for each condition to avoid familiarity and assignment of target words to the seven conditions was counterbalanced across participants.

Table 4.1: Experimental conditions and stimulus content.

	Stimuli	
Condition	Audio	Video
A	Male speaker	Black screen with grey fixation crosshair
V	None	Male speaker
AVc	Male speaker	Congruent male speaker
AVi	Male speaker	Incongruent male speaker
AVif	Male speaker	Incongruent female speaker[a]
AVin	Male speaker	Wildlife scenes with fixation crosshair
AVsf	Male speaker	Still image of male speaker’s face

[a] A different female speaker was used in each of the 15 trials to prevent association with the male speaker’s voice.

4.2.3 Behavioural Data Analysis

Participants’ performance on the target detection task was examined for multisensory effects. Specifically, it was examined whether reaction times (RTs) were facilitated by congruent bimodal speech (AVc) compared to unimodal speech (A, V), an effect known as a ‘redundant signals effect’ (RSE; Kinchla, 1974). An RSE does not necessarily imply multisensory interaction unless it violates the race model (Raab, 1962). The race model predicts that the RT in response to a bimodal stimulus is determined by the faster of the two unimodal processes. Violation of the race model was examined using the following inequality (Miller, 1982):

$$F_{\mathrm{AVc}}(t) = F_{\mathrm{A}}(t) + F_{\mathrm{V}}(t) - F_{\mathrm{A}}(t) \times F_{\mathrm{V}}(t), \tag{4.1}$$

where F_{AVc}, F_A and F_V are the cumulative distribution functions (CDFs) based on the RTs of the AVc, A and V conditions respectively. CDFs were generated for each participant and condition, divided into 9 quantiles (0.1, 0.2,…, 0.9) and group averaged (Ulrich et al., 2007).

4.2.4 EEG Acquisition and Pre-Processing

Continuous EEG data were acquired using an ActiveTwo system (BioSemi) from 128 scalp electrodes and two mastoid electrodes. The data were low-pass filtered online below 134 Hz and digitised at a rate of 512 Hz. Triggers indicating the start of each trial were recorded along with the EEG. These triggers were sent by an Arduino Uno microcontroller which detected an audio click at the start of each soundtrack by sampling the headphone output from the PC. Subsequent pre-processing was conducted offline in MATLAB; the data were band-pass filtered between 0.3 and 30 Hz, downsampled to 64 Hz and re-referenced to the average of the mastoid channels. To identify channels with excessive noise, the time series were visually inspected in Cartool (http://www.fbmlab.com/cartool-software/) and the standard deviation of each channel was compared with that of the surrounding channels in MATLAB. Channels contaminated by noise were recalculated by spline-interpolating the surrounding clean channels in EEGLAB (Delorme and Makeig, 2004). Trials contaminated by excessive low-frequency noise were detrended using a sinusoidal function in NoiseTools (http://audition.ens.fr/adc/NoiseTools/).

4.2.5 Stimulus Characterisation

Because the aim was to examine how visual information affects the neural tracking of auditory speech, the stimuli were characterised using the broadband envelope of the acoustic signal (Rosen, 1992). To model the frequency analysis of the auditory periphery (see section 2.1.2), the stimuli were first band-pass filtered into 128 logarithmically-spaced frequency bands between 100 and 6500 Hz using a gammatone filterbank (Yang et al., 1992). The upper- and lower-most filter limits captured the first, second and third formant spectral regions of the speech signals, known to carry the acoustic information that correlates most with visual speech features (Chandrasekaran et al., 2009, Grant and Seitz, 2000). The envelope in each of the 128 frequency bands was calculated using a Hilbert transform and the broadband envelope was obtained by averaging over the 128 narrowband envelopes.

4.2.6 Stimulus Reconstruction

To determine how faithfully the cortical activity tracked the speech envelope during each condition, an estimate of the speech envelope was reconstructed from the EEG data and compared to the original envelope (see Fig. 4.1). Stimulus reconstruction was implemented using the mTRF Toolbox described in the previous chapter (see section 3.2.4). The time lags were set between 0–500 ms, i.e., $\tau_{\text{min}} = -500$ and $\tau_{\text{max}} = 0$ samples (Ding and Simon, 2012b). Leave-one-out cross-validation was used to reconstruct an estimate of each of the 15 stimuli per condition. To optimise performance within each condition, a parameter search (over the range 2^{14}, 2^{15},…, 2^{21}) was conducted for the ridge value that maximised the correlation between the original and reconstructed envelopes. To prevent overfitting, the ridge parameter was tuned to the value that gave the highest mean reconstruction accuracy across the 15 trials (see section 3.3.2).

4.2.7 Quantifying Multisensory Interactions

The decision to include all 128 channels of EEG in the reconstruction analysis is justified because irrelevant filter channels can maintain zero weight whilst allowing the model to capture additional variance (Mesgarani et al., 2009, Pasley et al., 2012). However, this multi-channel approach required the application of different criteria when quantifying multisensory interactions in the congruent and incongruent AV conditions (see section 2.4.1). For the incongruent AV conditions (AVi, AVif, AVin, AVsf), a maximum criterion model approach was applied, i.e., each multisensory condition was compared to the optimal unisensory (A) condition. This was fair because the incongruent visual stimuli were not temporally correlated with the speech envelope; therefore, information encoded by the visual system in occipital channels did not benefit reconstruction of the envelope. However, this was not true for the congruent AV condition (AVc), where the dynamics of the visual stimulus were highly correlated with those of the speech envelope. This approach would allow the AVc model to infer complementary information from correlated visual speech processing as reflected on parieto-occipital channels (Luo et al., 2010, Bernstein and Liebenthal, 2014), even without ever explicitly quantifying those visual features in the model fitting. Previous work has attempted to circumvent this bias by restricting the analysis to only the frontal electrodes (Crosse et al., 2013). However, this approach significantly compounds model performance and, in any case, would not guarantee that the AVc condition was unbiased as volume conduction could still result in visual cortical activity being reflected in frontal channels.

Instead, multisensory interactions in the AVc condition was examined using the additive model criterion (Stein and Meredith, 1993, Barth et al., 1995, Berman, 1961). The rationale here is that multisensory interactions can be inferred from differences between cortical responses to multisensory stimuli and the summation of unisensory responses [i.e., AVc−(A+V); see section 2.4.1]. As previously mentioned, the validity of the additive model for the purpose of indexing multisensory integration in electrophysiological studies is well established (Besle et al., 2004b). The following procedure was used to apply the additive model approach to the stimulus reconstruction analysis:

1. New A and V reconstruction filters were calculated using the A and V data sets, respectively ($\lambda_A = 2^{14}, 2^{15},\ldots, 2^{20}$; $\lambda_V = 2^{14}, 2^{15},\ldots, 2^{34}$).
2. We calculated the algebraic sum of the A and V filters (A+V) for every combination of λ values.
3. Critically, each additive model was then assessed using the EEG data from the AVc condition – this ensured that the model could decode the envelope from channels that encoded both auditory and visual information.
4. A grid search was conducted to find the combination of λ values that maximised reconstruction accuracy across the 15 stimuli.

The difference between the AVc and A+V models was quantified in terms of how accurately each of them could reconstruct the speech envelopes from the AVc data using leave-one-out cross-validation. Such differences were interpreted as an index of multisensory integration. This multisensory cross-validation approach was implemented in mTRF Toolbox using the *mTRFmulticrossval* function (Appendix E).

4.2.8 Temporal Response Function Estimation

To visualise the temporal profile of the neural response to the different stimuli, the temporal response function at every channel was calculated (see Fig. 4.1). Unlike the stimulus reconstruction approach, it is not a multivariate regression, but represents multiple univariate mappings between stimulus and EEG (see section 3.2.1). TRF model parameters are neurophysiologically interpretable, i.e., non-zeros weights are only observed at channels where cortical activity is related to stimulus encoding (Haufe et al., 2014). This allows for examination of the amplitude, latency and scalp topography of the stimulus-EEG relationship, complementing the stimulus reconstruction approach. For each 60-s trial, the TRFs were calculated at time lags between −100 and 400 ms.

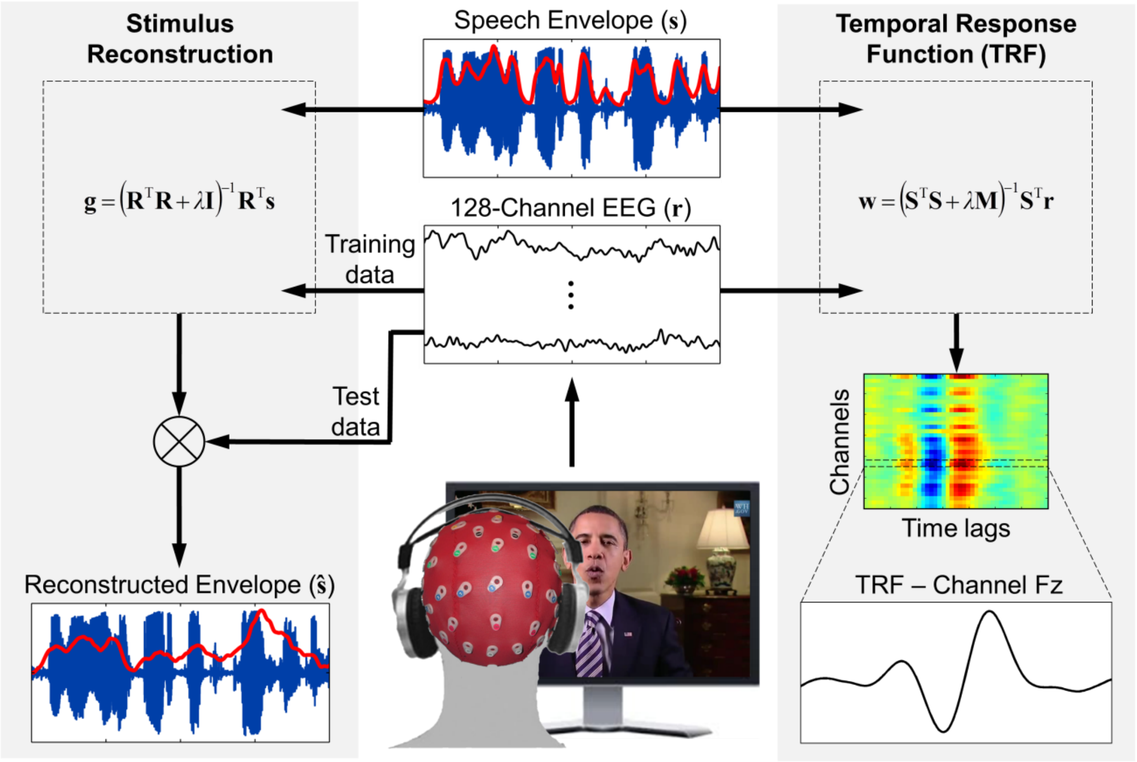

Stimulus Reconstruction
Speech Envelope (s)
Temporal Response Function (TRF)
$g = (R^T R + \lambda I)^{-1} R^T s$
128-Channel EEG (r)
$w = (S^T S + \lambda M)^{-1} S^T r$
Training data
Test data
Channels
Time lags
TRF – Channel Fz
Reconstructed Envelope (ŝ)

Figure 4.1: SI framework for studying natural AV speech processing.

Illustration of the working principle of the stimulus reconstruction and temporal response unction (TRF) approach implemented in mTRF Toolbox. The TRF approach is used for visualisation of the neural response dynamics, but not prediction because of its univariate limitations. The stimulus reconstruction approach is used to reconstruct the envelope, and model performance is interpreted as an index of envelope tracking. However, reconstruction model parameters are not neurophysiologically interpretable as in the TRF.

4.2.9 Multidimensional Scaling

In an effort to visualise any potentially interpretable differences between the various reconstruction models, non-metric multidimensional scaling (MDS) was applied to the model channel weights. MDS has been applied to electrophysiological data in previous studies to demonstrate the dissimilarity of neural responses elicited to different phonemes (Chang et al., 2010, Di Liberto et al., 2015). Given a set of objects, MDS works by embedding each object in a multi-dimensional space such that distances between objects produce an empirical matrix of dissimilarities. Here, the objects are the different stimulus conditions and the dissimilarities are the standardised Euclidean distances between the filter weights. To capture maximal model variance across the scalp, weight vectors from all 128 channels were concatenated and group averaged. To determine how many dimensions would be maximally required to explain model variance, Kruskall stress was measured as a function of dimensions (Kruskal and Wish, 1978). Two dimensions were sufficient to meet the criteria, i.e., stress < 0.1.

4.2.10 Statistical Analyses

Any effects of condition on behaviour or EEG measures were established using one-way repeated-measures ANOVAs, except where otherwise stated. Where sphericity was violated, the Greenhouse-Geisser corrected degrees of freedom are reported. *Post hoc* comparisons were conducted using two-tailed (paired) *t*-tests, except where one-tailed tests were necessary. Multiple pairwise comparisons were corrected for using the Holm-Bonferroni method. All numerical values are reported as mean ± SD.

4.3 Results

4.3.1 Behaviour

Twenty-one participants performed a target detection task during EEG recording. To examine whether the detection of auditory targets was affected by the visual stimulus, the reaction times and hit rates across the five AV conditions were compared (AVc, AVi, AVif, AVin, AVsf). The visual stimulus had a significant effect on RT ($F_{(4,80)}$ = 3.13, p = 0.02) but not on hit rate, which was near ceiling (median > 92%; $\chi^2_{(4)}$ = 7.49, p = 0.11; Friedman test). To test for an RSE, planned *post hoc* comparisons were made between the congruent AV condition (AVc) and the unimodal conditions (A, V; Fig. 4.2*A*). RTs for the AVc condition (586 ± 92 ms) were significantly faster than those for both the A condition (620 ± 88 ms; $t_{(20)}$ = 2.74, p = 0.01) and the V condition (819 ± 136 ms; $t_{(20)}$ = 7.9, p = 1.4×10^{-7}), confirming an RSE. To test whether this RSE exceeded the statistical facilitation predicted by the race model, we compared the bimodal CDFs with the sum of the unimodal CDFs (Fig. 4.2*B*). Three participants were excluded from this analysis as they did not detect enough targets in the V condition to allow estimation of the CDF. The race model was violated by > 50% of participants at the first two quantiles but the effect was not significant (first quantile: $t_{(17)}$ = 0.01, p =

0.5; second quantile: $t_{(17)} = 0.16$, $p = 0.56$; one-tailed tests). This is likely due to the nature of the task involving, as it did, an easy auditory detection task and much more difficult visual detection (lipreading) task. As such, RTs in the AVc condition were likely dominated by reaction to the auditory stimulus with minimal contribution from the visual modality. None of the incongruent AV conditions (AVi, AVif, AVin, AVsf) showed behavioural differences relative to the A condition or each other.

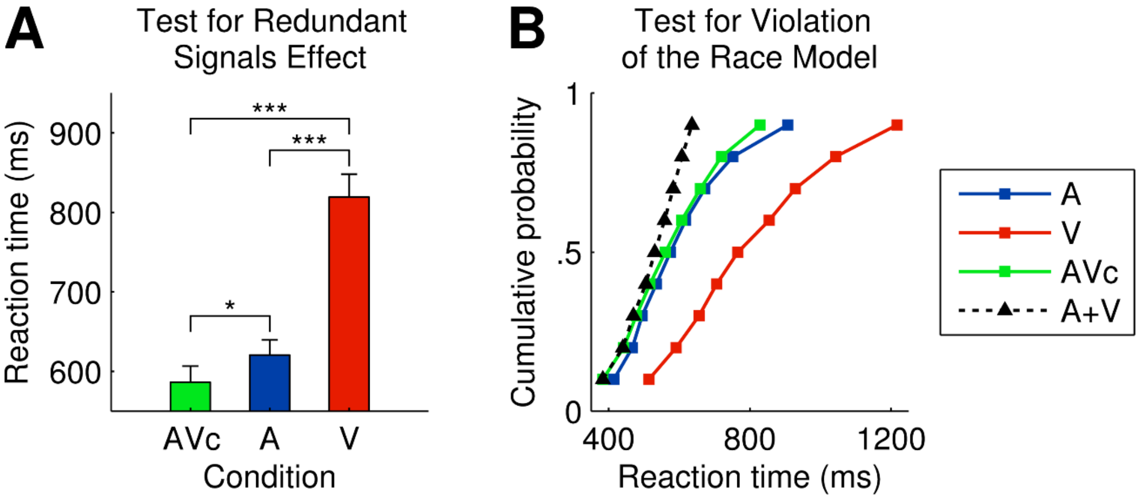


Figure 4.2: Examination of behaviour under the race model.
A, Mean ($N = 21$) reaction times for the congruent audiovisual (AVc; green), audio-only (A; blue) and visual-only (V; red) condition. Error bars indicate SEM across participants. Brackets indicate pairwise statistical comparisons (*$p < 0.05$, ***$p < 0.001$). ***B***, Group-average ($N = 18$) cumulative distribution functions based on the reaction times shown in ***A***. The dashed black trace represents the facilitation predicted by the race model (A+V).

4.3.2 Impact of AV Congruency on Envelope Tracking

To investigate the impact of AV congruency on the cortical representation of speech, an estimate of the speech envelope from the EEG data was reconstructed for each condition (Fig. 4.3*A*). Critically, it was found that the envelope was encoded more accurately by congruent AV speech (AVc; Pearson's $r = 0.2 \pm 0.05$) than could be explained by the additive model (A+V; 0.18 ± 0.04; $t_{(20)} = 3.84$, $p = 0.001$; Fig. 4.3*B*). This suggests that, even in optimal listening conditions, congruent visual speech enhances neural tracking of the acoustic envelope in line with my primary hypothesis.

As discussed above, quantifying multisensory interactions in the incongruent AV conditions (AVi, AVif, AVin, AVsf) simply involved direct comparisons with the

A condition. Across these five conditions, there was a significant effect of visual stimulus on reconstruction accuracy ($F_{(2,40.3)}$ = 11.84, p = 8.8×10^{-5}; Fig. 4.3*B*). However, *post hoc* comparisons revealed that envelope tracking was not enhanced by incongruent AV speech relative to unimodal speech. This suggests that the neural mechanism underlying enhanced envelope tracking in the case of congruent AV speech relies on discrete, phasic interactions as opposed an ongoing, tonic process; in other words, it is likely that the temporal coherence between auditory and visual speech is of paramount importance to this multisensory enhancement. Although an enhancement effect was not observed, the data indicate that envelope tracking was actually inhibited by incongruent AV speech, but only when the visual stimulus was incongruent both temporally and contextually. Relative to the A condition (0.17 ± 0.05), envelope tracking was significantly inhibited by the presentation of an incongruent female speaker (AVif; 0.15 ± 0.05; $t_{(20)}$ = 3.3, p = 0.004) and incongruent nature scenes (AVin; 0.16 ± 0.05; $t_{(20)}$ = 2.3, p = 0.03). Unsurprisingly, reconstruction accuracy (of the acoustic envelope) was lowest in the V condition (0.13 ± 0.04), yet, it maintained accuracy significantly above the 95th percentile of chance level (shaded area, Fig. 4.3*B*). This demonstrates the efficacy of the stimulus reconstruction method to infer temporally correlated information pertaining to one sensory modality from another.

Recently, Ding and Simon (2013) demonstrated that the accuracy with which the envelope can be reconstructed from MEG data is highly correlated with stimulus intelligibility across participants. This could only be demonstrated at a signal-to-noise ratio (SNR) where intelligibility scores were at an intermediate level, i.e., ~50%. In the present chapter, the V condition was the only one where hit rate was not at ceiling (36.8 ± 18.1%). Under the assumption that hit rate is also reflective of intelligibility, the correlation coefficient between each participant’s mean reconstruction accuracy and hit rate was calculated using the V data (Fig. 4.3*C*).This measure of behaviour was significantly correlated with reconstruction accuracy across participants (r = 0.7, p = 6.6×10^{-4}). Participant 13 was excluded from this analysis as an outlier as they reported an inability to detect any targets during the V condition (indicated by ‘×’ marker in Fig. 4.3*C*).

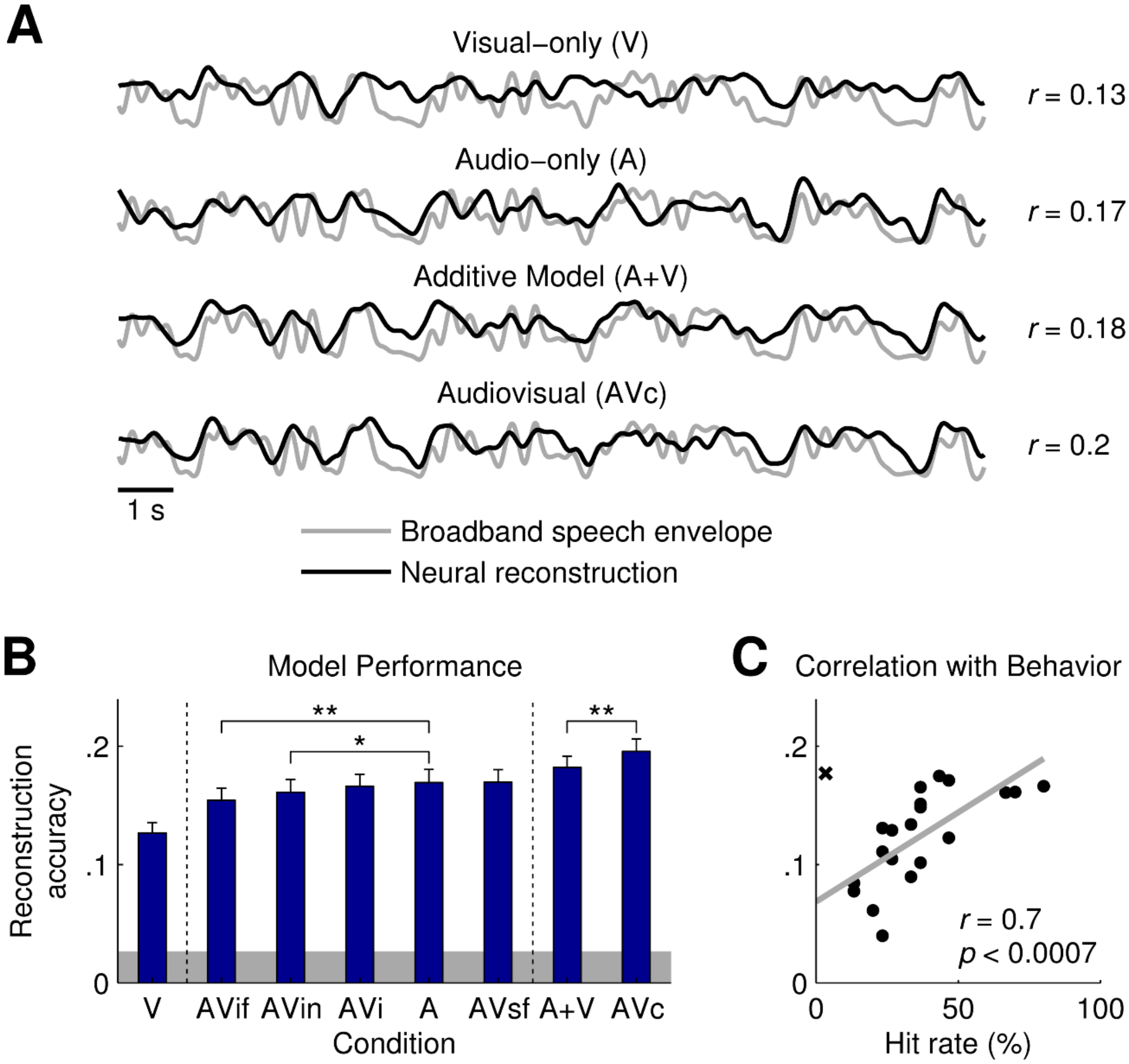


Figure 4.3: Reconstruction of the speech envelope from EEG.

A, Examples of the original speech envelope (grey) with the group-average neural reconstruction (black) superimposed. Signals were filtered below 3 Hz for visualization. The mean correlation coefficient between the original and reconstructed envelopes (i.e., reconstruction accuracy) is shown to the right. ***B***, Mean (N = 21) reconstruction accuracy for all eight models in ascending order. Error bars indicate SEM across participants. Dashed lines indicate planned *post hoc* sub-groups and brackets indicate pairwise statistical comparisons (*$p < 0.05$, **$p < 0.01$). The shaded area represents the 95th percentile of chance level reconstruction accuracy (permutation test). ***C***, Correlation (N = 20) between reconstruction accuracy and hit rate using visual speech data. Each datapoint represents a participant's mean value and the '×' marker indicates the participant that was excluded from the analysis. The grey line represents a linear fit to the data.

4.3.3 Temporal Scale of AV Speech Integration

It has been suggested that AV speech perception includes the neuronal integration of temporally fine-grained correlations between the auditory and visual speech stimuli, even at the phonetic level (Grant and Seitz, 2000, Klucharev et al., 2003). In contrast, other work has suggested that, at least in some detection paradigms, neuronal integration at this detailed level of granularity is not necessary (Tjan et al., 2014). The current chapter aimed to elucidate whether the multisensory effects [i.e., AVc > (A+V)] may be occurring on the timescale of phonemes, syllables, words or sentences. To do this, the correlation coefficients between the reconstructed and original envelopes at every 2-Hz-wide frequency band between 0 and 30 Hz were calculated. Figure 4.4*A* shows reconstruction accuracy as a function of frequency for the AVc and A+V models, while Fig. 4.4*B* shows the multisensory interaction effect [AV−(A+V)] at each frequency band. Significant multisensory effects were measured at 2–4 Hz ($t_{(20)} = 4.74$, $p = 1.3\times10^{-4}$) and 4–6 Hz ($t_{(20)} = 4.1$, $p = 5.6\times10^{-4}$). This suggests that neural tracking of the acoustic envelope is enhanced by congruent visual speech at a temporal scale that corresponds to the rate of syllables. There was also a significant effect at 16–18 Hz ($t_{(20)} = 3.8$, $p = 0.001$), although this finding is less compelling given the low reconstruction SNR at this frequency range.

A related question is whether temporal scales that are optimal for reconstructing the acoustic envelope from visual speech data can be ascertained. Addressing this issue is not entirely straightforward because there are many visual speech features at different levels of temporal granularity that correlate with the acoustic envelope (Jiang et al., 2002, Jiang and Bernstein, 2011, Chandrasekaran et al., 2009). In the stimulus reconstruction approach, the model reflects not only activity from auditory cortex that tracks the dynamics of the acoustic envelope, but also activity from potentially any visual area whose activity is correlated with the acoustic envelope and reflected in the EEG (Luo et al., 2010). Indeed, the reconstruction model can also indirectly index activity in brain areas whose activity is correlated with the acoustic envelope, even if that activity is not reflected directly in the data (Mesgarani et al., 2009). In one way, this is an advantage of the approach in that it is sensitive to visual speech processing without having to explicitly define specific visual speech features. However, it also makes it very difficult to tease apart the details of those visual speech contributions.

Bearing this in mind, we examined which frequencies optimised reconstruction of the acoustic envelope from the V data and compared it to those that optimised reconstruction using the A data (Fig. 4.4*C*). Reconstruction accuracy was significantly higher in the A condition at almost every frequency band (p < 0.05, t-tests; Holm-Bonferroni corrected; Fig. 4.4*D*) except at two distinct spectral regions which, interestingly, corresponded to the two peaks in multisensory enhancement (2–4 Hz: $t_{(20)}$ = 1.8, p = 0.08; 16–18 Hz: $t_{(20)}$ = 0.17, p = 0.87).

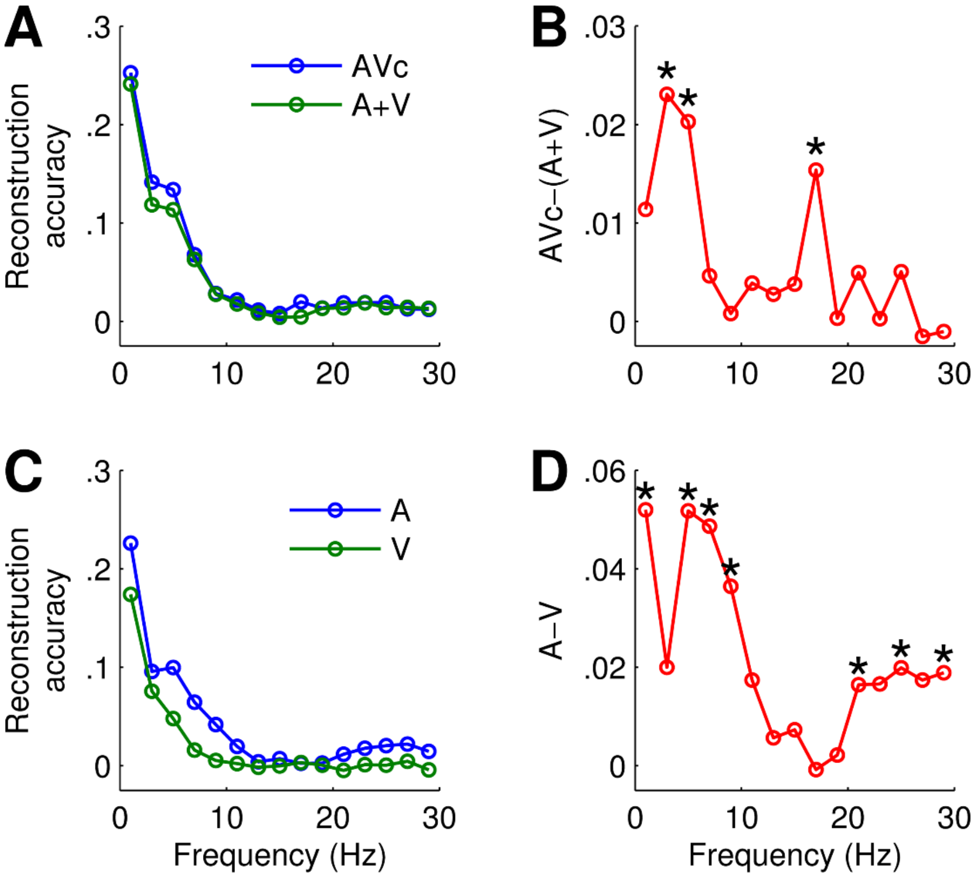


Figure 4.4: Reconstruction of the speech envelope at different temporal scales.
A, Mean (N = 21) reconstruction accuracy as a function of envelope frequency for the congruent audiovisual (AVc; blue) and additive (A+V; green) models. ***B***, Multisensory interaction effect [AVc−(A+V)] at each frequency band (*p < 0.05, t-tests; Holm-Bonferroni corrected). ***C***, Mean (N = 21) reconstruction accuracy as a function of envelope frequency for the audio-only (A; blue) and visual-only (V; green) models. ***D***, Differences in unimodal model performance (A−V) at each frequency band (*p < 0.05, t-tests; Holm-Bonferroni corrected).

4.3.4 Spatiotemporal Representation of AV Speech

To examine the temporal profile of our neuronal multisensory effects, the temporal response function for each of the seven conditions was determined, as well as the sum of the unimodal TRFs (A+V). Figure 4.5*A* shows the temporal profile of the TRFs for the congruent speech conditions at frontal channel Fz (top) and occipital channel Oz (bottom), while Fig. 4.5*B* shows the TRFs for the incongruent speech conditions at the same channel locations. Comparing AVc with A+V as before, we see multisensory interaction effects in the form of a reduced amplitude over occipital scalp at ~140 ms (Oz: $t_{(20)} = 2.9$, $p = 0.01$; Fig. 4.5*C*, top) and over frontal scalp at ~220 ms (Fz: $t_{(20)} = 3.1$, $p = 0.006$; Fig. 4.5*C*, bottom). The TRFs for the incongruent speech conditions were all identical to that of the A condition except for AVin, where the $\mathrm{P2}_{\mathrm{TRF}}$ component was significantly reduced in amplitude at several frontocentral electrode sites ($p < 0.05$; Holm-Bonferroni corrected).

To relate our late neuronal multisensory effect back to the stimulus reconstruction results, the relative channel weightings of each of the reconstruction models were examined. The channel weights represent the amount of information that each channel provides for reconstruction, i.e., highly informative channels receive weights of greater magnitude while channels providing little or no information receive weights closer to zero. However, unlike TRF model parameters, significant non-zero weights may also be observed at channels where cortical activity is statistically independent of stimulus tracking, hence the spatiotemporal distribution of such model weights can be difficult to interpret in terms of underlying neural generators (Ding and Simon, 2012a, Haufe et al., 2014).

Figure 4.5*D* shows the channel weighting for each model averaged over time lags that correspond to the neuronal multisensory effects (125–250 ms). Although not necessarily reflective of the underlying neural generators, the model weights clearly maintain distinct topographic patterns subject to stimulus modality. Channels over left and right temporal scalp make large contributions to stimulus reconstruction in the A model, while channels over occipital scalp are dominant in the V model. Unsurprisingly, channels over both temporal and occipital scalp both make significant contributions in the congruent AV model, while only channels over temporal scalp make significant contributions in the incongruent AV models. This is because the incongruent visual stimuli were not informative of the acoustic envelope dynamics. The

A+V model places significant weight on channels over temporal and occipital scalp, similar to the AVc model.

While the AVc and A+V models appeared to have similar channel weightings, their ability to decode the speech envelope was significantly different. To better visualise the similarity relationships across all eight models, the channel weight dissimilarity in a two-dimensional Euclidean space were represented using non-metric MDS. Model dissimilarity was examined within two specific time intervals; an early interval (0–125 ms; Fig. 4.5*E*, left), at which latencies there were no multisensory effects evident in our TRF measures, and a later interval (125–250 ms; Fig. 4.5*E*, right), at which latencies there were significant multisensory effects evident in the TRFs. Visual inspection of the MDS plot for the earlier time interval (Fig. 4.5*E*, left) suggests that the models were organised into two discrete groupings consisting of audio and non-audio stimuli. The AVc model is not visually discriminable from the other audio conditions at this interval, in line with the TRFs. In the later interval however (Fig. 4.5*E*, right), the AVc model shows the greatest discriminability relative to the other models, indicating that it is capturing neuronal contributions from crossmodal interactions that are not well represented in the A+V model, also in agreement with the TRF results.

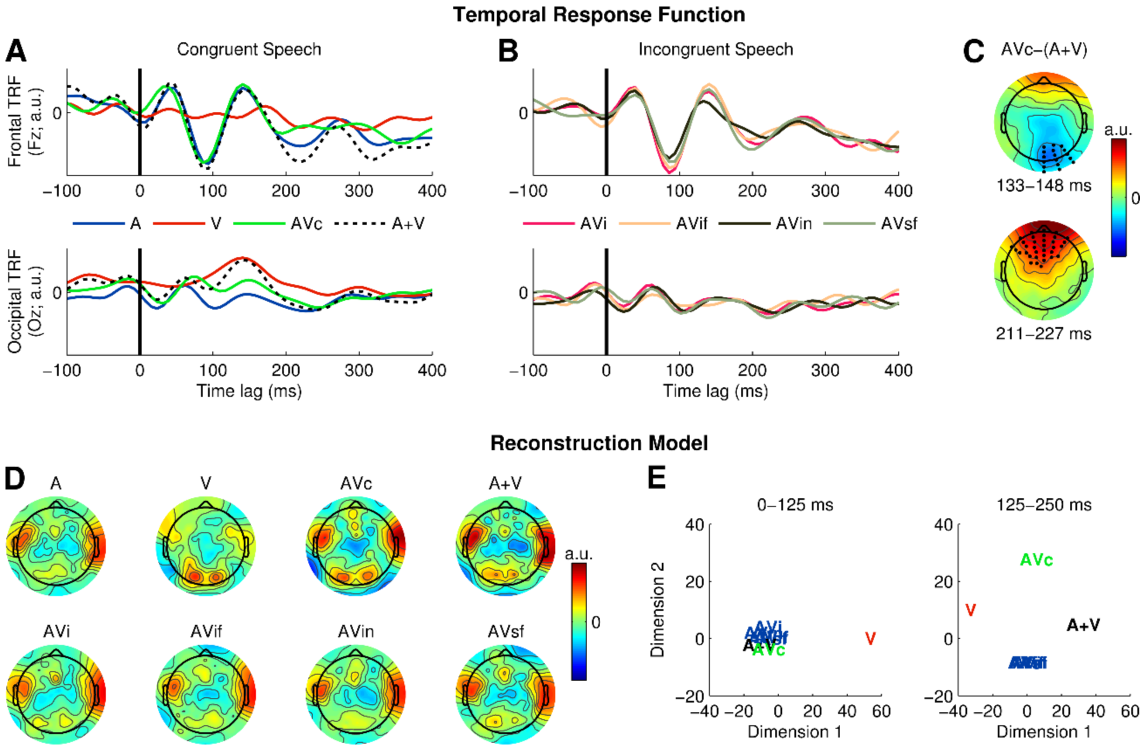


Figure 4.5: Spatiotemporal analysis of neuronal multisensory effects.

A, Group-average (*N* = 21) temporal response functions (TRFs) for congruent speech conditions at frontal scalp location Fz (top) and occipital scalp location Oz (bottom). ***B***, TRFs for incongruent speech conditions at the same scalp locations as in ***A***. ***C***, Topographic maps of multisensory interaction effects [AV−(A+V)] at ~140 ms (top) and ~220 ms (bottom). The black markers indicate channels where the multisensory effect was significant across participants ($p < 0.05$, *t*-tests). ***D***, Group-average (*N* = 21) reconstruction models, highlighting differential channel weightings at time lags corresponding to neuronal multisensory effects in ***C*** (125–250 ms). ***E***, Visualization of filter weight dissimilarity in a two-dimensional Euclidean space obtained using non-metric multidimensional scaling for time lags between 0–125 ms (left) and 125–250 ms (right). Colouring was applied to highlight discrete groupings based on the 125–250 ms interval.

4.4 Discussion

In this chapter, it has been demonstrated that when visual speech is congruent with auditory speech, the cortical representation of the speech envelope is enhanced relative to that predicted by the additive model criterion. These crossmodal interactions were most prominent at timescales indicative of syllabic integration (2–6 Hz). This was reflected in the neural responses by a suppression in amplitude at ~140 ms and ~220 ms which corresponded with a late shift in the spatiotemporal profile of our reconstruction models, suggesting the involvement of neural generators that were not strongly activated during unimodal speech.

4.4.1 Congruent Visual Cues and Envelope Tracking

The enhancement in cortical entrainment produced by AVc speech exceeded that predicted by the additive model. This fits with recent views on AV speech processing which suggest that visual speech increases the accuracy with which auditory cortex tracks the ongoing speech signal, leading to improved speech perception (Peelle and Sommers, 2015, Schroeder et al., 2008). However, we cannot rule out the possibility that attention was enhanced by AV stimulation (Talsma et al., 2010), and the fact that enhanced attention leads to more accurate envelope tracking (O'Sullivan et al., 2015, Ding and Simon, 2012a). In contrast, a recent MEG study did not demonstrate enhanced neural tracking for single-speaker AV speech, but did for competing speakers (Zion-

Golumbic et al., 2013a). However, their finding was based on inter-trial coherence, an indirect measure of envelope tracking, whereas stimulus reconstruction and TRF estimation are direct measures and, as such, may be more sensitive to subtle differences in tracking elicited during single-speaker AV speech. In support of this, they did in fact report enhanced TRF amplitude for single-speaker AV speech compared to single-speaker A speech (Zion-Golumbic et al., 2013a). Furthermore, their stimuli were shorter (~10 s) and were repeated more times (40 per condition), meaning that the contribution of the visual stimulus may have varied based on the ability of participants to predict the upcoming auditory information.

Indeed, the effects of being able to predict the acoustic information may also be reflected by the fact that TRF amplitude was enhanced as early as ~50 ms (Zion-Golumbic et al., 2013a). In contrast, the present results suggest that TRF amplitude was reduced at the later latencies of ~140 ms and ~220 ms. This finding fits with numerous ERP/ERF studies that have demonstrated late multisensory interactions in the form of subadditive cortical measures between 100–250 ms (Besle et al., 2004a, Bernstein et al., 2008, van Wassenhove et al., 2005, Stekelenburg and Vroomen, 2007, Arnal et al., 2009, Möttönen et al., 2004). Such deactivation effects have been linked to several theories such as predictive coding (Arnal et al., 2009, Arnal et al., 2011, van Wassenhove et al., 2005), cross-sensory inhibition and dedication of attentional resources to the relevant modality (Besle et al., 2004b). However, in keeping with recent perspectives on AV speech processing (Peelle and Sommers, 2015), it is postulated that this late suppression of cortical activity is reflective of an emergent integration stage that utilises the relevant visual speech information to constrain the number of possible candidates. Indeed, this notion that emergent neuronal contributions may be driving our multisensory effects was also supported by our MDS analysis of the reconstruction models which revealed differential AVc versus A+V weight patterns only at later time lags (125–250 ms). It has been suggested (Peelle and Sommers, 2015) that earlier integration effects are likely reflective of increased auditory cortical sensitivity to acoustic information, hence we predict that they may be more evident in complementary modes of AV speech such as speech-in-noise.

4.4.2 Incongruent Visual Cues and Envelope Tracking

While envelope tracking was shown to be enhanced by congruent AV speech, it was inhibited when the auditory and visual streams were both temporally and contextually incongruent. This fits with seminal fMRI work which investigated multisensory integration in continuous passages (~30 s) of natural speech (Calvert et al., 2000). Specifically, they found that congruent AV speech elicited superadditive activation in pSTS, whereas incongruent AV speech elicited subadditive activation. It is therefore possible that the effects observed in the envelope tracking measures presented here may reflect contributions from pSTS. Whether or not this is specific to incongruent speech remains unclear. Related to this, a recent MEG study which used congruent and incongruent naturalistic AV videos demonstrated increased inter-trial coherence for congruent stimuli relative to incongruent stimuli (Luo et al., 2010). However, as they did not examine unisensory speech, it is difficult to determine whether the incongruent stimuli actually inhibited neural tracking or just failed to enhance it.

Another possible explanation relates to the role of attention during incongruent stimulation. In the AVin condition, the male speaker's voice is less relevant to the visual stimulus (nature scenes), hence attentional resources dedicated to the auditory stimulus may have been reduced, a situation that is known to impact upon speech tracking (O'Sullivan et al., 2015, Ding and Simon, 2012a). This is further supported by the decrease in $P2_{TRF}$ amplitude (AVin), an effect that has also been linked to reduced attention (Power et al., 2012). This notion also fits with the theory that during conflicting AV presentation such as the McGurk scenario, directing attention towards a particular modality tends to reduce the bias of the unattended modality (Welch and Warren, 1980, Talsma et al., 2010). In the AVif condition, attention may have been modulated in a slightly different way. According to the “attention in time” hypothesis (Large and Jones, 1999, Jones et al., 2006, Nobre et al., 2007, Nobre and Coull, 2010), entrainment may have been less effective because attention was being directed towards the auditory stream at time points that were acoustically less relevant. However, if this were the case, then one would expect to see a reduction in envelope tracking for the incongruent male condition (AVi) as well. Given that context seemed to play an important role in this inhibitory effect, the former explanation seems the more likely.

4.4.3 AV Speech Integration at the Syllabic Timescale

The present data suggest that envelope tracking is enhanced by congruent visual speech at a timescale commensurate with the rate of syllables (2–6 Hz). This fits very well with work by Luo et al. (2010) which demonstrated that the phase of auditory cortex tracks both auditory and visual stimulus dynamics and that this cross-modal phase modulation is most prominent in low-frequency neural information in the delta-theta band (2–7 Hz). This also fits with recent data that demonstrated a temporal correspondence between facial movements and the speech envelope in the 2–7 Hz frequency range (Chandrasekaran et al., 2009). Interestingly, there was no significant difference in the contribution from visual and auditory speech at frequencies where multisensory integration peaked. This may suggest that multisensory integration is enhanced for temporal scales where neither modality is particularly dominant, or at least where visual speech provides complementary information.

Future paradigms involving manipulations to the SNR of both the acoustic signal (e.g., speech-in-noise) and the visual signal (e.g., use of point light stimuli, dynamic annulus stimuli, partially occluded faces) may lead to shifts in the spectral profile of the multisensory effects and/or the unisensory effects, allowing firmer conclusions to be drawn (see Chapter 4 for examination of AV speech-in-noise). This endeavour might be aided further by extending the framework in order to reduce the reliance on the acoustic envelope by directly incorporating information about phonemes and visemes, as has been done recently for auditory speech research (Di Liberto et al., 2015). In addition, utilizing other approaches to quantify AV correlations such as those based on mutual information models (Nock et al., 2002) and Hidden Markov Models (Rabiner, 1989) may provide important complementary insights.

Chapter 5 Eye Can Hear Clearly Now: Inverse Effectiveness in Natural Audiovisual Speech Processing Relies on Long-Term Crossmodal Temporal Integration

5.1 Introduction

In the previous chapter, detection accuracy was not enhanced by the presentation of AV speech because performance in audio-only speech was already at ceiling. Indeed, the behavioural benefits of AV speech are more apparent in acoustic conditions where intelligibility is reduced (Bernstein et al., 2004b, Sumby and Pollack, 1954, Grant and Seitz, 2000, Erber, 1975, Ross et al., 2007a). Enhanced multisensory processing in response to weaker sensory inputs is a phenomenon known as inverse effectiveness (Meredith and Stein, 1986a). However, in the context of AV speech processing, there are particular audio signal-to-noise ratios (SNRs) at which the benefits of multisensory processing become maximized – a sort of multisensory 'sweet spot' (see section 2.4.2; Ross et al., 2007a). It is likely that when processing AV speech in such conditions, the brain must exploit both correlated and complementary visual information in order to optimize intelligibility (Campbell, 2008, Summerfield, 1987, Grant and Seitz, 2000). This could be achieved through multiple integration mechanisms, occurring at different temporal stages (see section 2.4.4). Specifically, recent perspectives on multistage AV speech processing suggest that visual speech provides cues to the timing of the acoustic signal that could project directly from visual cortex, increasing the sensitivity of auditory cortex to the upcoming acoustic information, while complementary visual cues that convey place and manner of articulation could be integrated with converging acoustic information in supramodal regions such as superior temporal sulcus (STS), serving to constrain lexical selection (see Peelle and Sommers, 2015).

Indeed, studying how the brain utilizes the timing and lexical constraints of visual speech to enhance the processing of acoustic information necessitates the use of natural, conversation-like speech stimuli. Recent electroencephalography (EEG) and magnetoencephalography (MEG) studies have used naturalistic speech stimuli to examine how visual speech effects the cortical representation of the speech envelope (Crosse et al., 2015a, Zion-Golumbic et al., 2013a). However, it is not yet known how these neural measures of speech processing are affected by visual speech at much lower SNRs where the multisensory processing is optimized. In particular, the specific neural mechanisms invoked in such situations are poorly understood. A recent MEG study examined how different levels of noise affect the cortical representation of audio-only speech and demonstrated that it is relatively insensitive to background noise, even at low SNRs where intelligibility is diminished (Ding and Simon, 2013). Only when intelligibility reached peri-threshold level (e.g., at an SNR of −9 dB), did they find that envelope tracking was significantly reduced. Given that AV speech has been shown to improve intelligibility in noise equivalent to an increase in SNR of up to 15 dB (Sumby and Pollack, 1954), we hypothesized that the addition of visual cues could substantially restore envelope tracking in such peri-threshold conditions.

In this chapter, an AV speech-in-noise paradigm was implemented to study the neural interaction between continuous auditory and visual speech at an SNR where multisensory processing was of maximal benefit relative to unisensory processing. As before, high-density EEG recordings were analysed using the multisensory framework introduced in Chapter 4. This chapter provides clear evidence that neural entrainment to continuous AV speech conforms to the principle of inverse effectiveness, and that it does so specifically by restoring early tracking of the speech signal and integrating low-frequency crossmodal information over longer temporal windows. These findings support the notion that different integration mechanisms contribute to AV speech processing over multiple stages (Peelle and Sommers, 2015, van Wassenhove et al., 2005, Eskelund et al., 2011, Baart et al., 2014, Schwartz et al., 2004). Our results also suggest that in degraded listening environments, crossmodal integration of AV speech occurs at a more coarse-grained linguistic level. The results of this study were presented at the *16th International Multisensory Research Forum* in Pisa in June, 2015 (Appendix

H) and at the *45th Annual meeting of the Society for Neuroscience* in Chicago in October, 2015 (Appendix I) and have been accepted for publication in *The Journal of Neuroscience*.

5.2 Methods

In order to examine how AV speech processing is affected by SNR, some of the data from Chapter 4 were re-analysed along with new ‘speech-in-noise’ data. Both of the experiments employed the same target detection task but involved separate participant samples.

5.2.1 Participants

Twenty-one participants (6 females; age range: 21–35 years) completed the speech-in-noise experiment. All participants were native English speakers, had self-reported normal hearing and normal or corrected-to-normal vision, were free of neurological diseases and provided written informed consent. All procedures were undertaken in accordance with the Declaration of Helsinki and approved by the Ethics Committee of the Health Sciences Faculty at Trinity College Dublin. For details of those that participated in the speech-in-quiet experiment, please refer to Chapter 4 (see section 4.2.1).

5.2.2 Stimuli and Procedure

The stimuli used in speech-in-noise experiment were the same videos used in the audio-only (A), visual-only (V) and congruent audiovisual (AV) conditions in the speech-in-quiet experiment. Please refer to Chapter 4 for exact video specifications (see section 4.2.2). For the speech-in-noise experiment, the video soundtracks were mixed with spectrally-matched stationary noise to ensure consistent masking across stimuli (see Appendix A, Fig. A1; Ding and Simon, 2013, Ding et al., 2014). 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 (48 kHz; Parsons, 1987).

Behavioural piloting was used to select the SNR value (as measured by RMS) such that it maximised the increase in intelligibility produced by AV speech relative to A speech (i.e., inverse effectiveness). A subset of participants (*N* = 3) listened to four

60-s passages of A and AV speech at SNRs of −7, −9 and −11 dB. After each passage, they were asked to rate in percent how intelligible the speech was. These data indicated that an SNR of −9 dB yielded the largest perceptual gain (see Fig. 5.1) and thus was chosen for the main experiment. Note that SNR is a relatively unstable measure because it is highly dependent on certain characteristics of the speech material such as dynamic range and the prevalence of gaps. Therefore, instead of choosing the SNR value based on values reported in previous studies, it is better to identify the part of the psychometric function of interest and then work backwards. The same spectrally-matched noise stimuli were also presented in the V condition, but without any speech content.

Task instructions and testing conditions in the speech-in-noise experiment were identical to those described in Chapter 4 (see section 4.2.2). In addition to detecting target words, participants in the speech-in-noise experiment were required to subjectively rate 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 ten-point scale (0–10%, 10–20%,…, 90–100%). While stimulus presentation order was completely random in the speech-in-quiet experiment, this approach was not suitable for the speech-in-noise paradigm, because if the same speech passage was presented twice in quick succession (albeit in different conditions), it could potentially influence intelligibility in the latter condition. Instead, the stimuli were presented such that a particular speech passage could not be repeated in another condition within 15 trials of the preceding one. Thus, the 15 passages were presented in the same order three times but the condition from trial-to-trial was randomised.

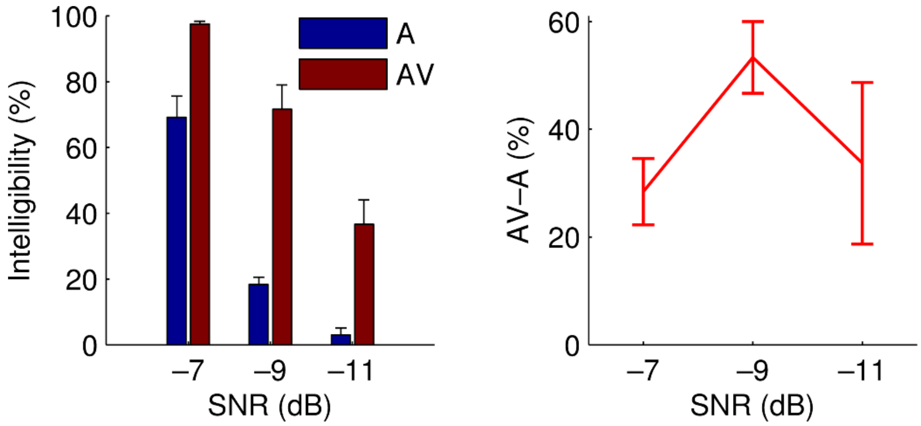


Figure 5.1: Pilot data used to identify multisensory ‘sweet spot’.

The left panel shows the mean subjectively-rated intelligibility for audio-only (A) and audiovisual (AV) speech at each SNR and the right panel shows the multisensory gain (AV−A) as a function of SNR. Error bars indicate SEM.

5.2.3 Behavioural Data Analysis

To identify a behavioural measure of multisensory integration (MSI), we examined whether the probability of detecting a multisensory stimulus exceeded the statistical facilitation produced by the unisensory stimuli. False positives were accounted for by taking an F-measure of each participant's detection rate. F-scores (or F1 scores) were calculated as the harmonic mean of precision and recall (Rijsbergen and Joost, 1979). Thus, our behavioural MSI measure was calculated as follows:

$$\mathrm{MSI}_{\mathrm{Behav}} = F_1(\mathrm{AV}) - \hat{F}_1(\mathrm{AV}) \qquad (5.1)$$

where $F_1(\mathrm{AV})$ is the F1 score for the AV condition and $\hat{F}_1(\mathrm{AV})$ is the predicted F1 score based on the values of the unisensory conditions. Although the same detection task was implemented in both experiments, two different criteria were used to quantify $\hat{F}_1(\mathrm{AV})$ as outlined in Stevenson et al. (2014a). For speech-in-quiet, detection accuracy was near ceiling so a maximum criterion model was used: $\hat{F}_1(\mathrm{AV}) = \max[F_1(\mathrm{A}), F_1(\mathrm{V})]$. For speech-in-noise, accuracy was not at ceiling so a more conservative model was used that accounted for statistical facilitation (Blamey et al., 1989): $\hat{F}_1(\mathrm{AV}) = F_1(\mathrm{A}) + F_1(\mathrm{V}) - F_1(\mathrm{A}) \times F_1(\mathrm{V})$. Essentially, the term on the right represents the detection rate that would be expected when auditory and visual stimuli were presented together and processed independently (Stevenson et al., 2014a). To quantify the gain in performance produced by AV speech, we calculated $\mathrm{MSI}_{\mathrm{Behav}}$ as a percentage of $\hat{F}_1(\mathrm{AV})$, in other words, as a percentage of independent unisensory processing (see section 2.2.1 for further details).

5.2.4 EEG Acquisition and Pre-Processing

EEG data were acquired using the same high-density recording system and experimental protocol described in section 4.2.4. Pre-processing conducted offline, including filtering and artefact rejection, was identical to that described in section 4.2.4.

5.2.5 Stimulus Characterisation

In this study, EEG analysis focused on the speech signal below 3 kHz because the strongest correlation between the mouth opening and vocal acoustics is between 2–3 kHz (Chandrasekaran et al., 2009, Grant and Seitz, 2000, Grant, 2001), meaning that visual speech provides cues to the timing of less salient auditory events within this frequency range (see section 2.3.1). Furthermore, visual speech can offer complementary information in the form of place of articulation, which can help distinguish ambiguous acoustic content in second formant space.

The spectrogram representation of each stimulus was generated using a compressive gammachirp auditory filterbank that modelled the auditory periphery (Irino and Patterson, 2006). Outer and middle ear correction were applied using an FIR minimum phase filter before the stimuli were band-pass filtered into 256 logarithmically-spaced frequency bands between 80 and 3000 Hz. 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.

The rates of different linguistic units (e.g., words, syllables, vowels, consonants) in the speech stimuli were extracted from the audio files using the Forced Alignment and Vowel Extraction (FAVE) Software Suite (Rosenfelder et al., 2011). This returns the start and end time-points for individual phonemes, enabling detailed characterization of the timescale of both segmental and suprasegmental speech units.

5.2.6 Stimulus Reconstruction

Neural tracking of the speech signal was measured using the stimulus reconstruction technique described in Chapter 3 (see section 3.2.4). Decoders were optimised using the same specifications described in Chapter 4 (see section 4.2.6). The objective, still, was to reconstruct the underlying speech envelope (as opposed to the actual speech-in-noise mixture) because we only care about how the brain processes speech information. In any case, previous work has demonstrated that the underlying speech signal can be reconstructed from cortical activity with greater accuracy than the actual speech-in-noise mixture (Ding and Simon, 2013). As with the behavioural data, we define a neural measure of multisensory integration:

$$\mathrm{MSI}_{\mathrm{EEG}} = \mathrm{corr}\left[\hat{s}_{\mathrm{AV}}(t), s(t)\right] - \mathrm{corr}\left[\hat{s}_{\mathrm{A+V}}(t), s(t)\right] \quad (5.2)$$

where $\hat{s}_{\mathrm{AV}}(t)$ is the reconstructed envelope for the AV condition and $\hat{s}_{\mathrm{A+V}}(t)$ is the estimated envelope for the additive unisensory model (see section 4.2.6). Similar to the behavioural analysis, we defined multisensory gain by calculating MSI$_{\mathrm{EEG}}$ as a percentage of $\mathrm{corr}\left[\hat{s}_{\mathrm{A+V}}(t), s(t)\right]$ (see section 2.4.1 for further details).

5.2.7 Single-lag Reconstruction Analysis

When reconstructing the speech envelope, the decoder $g(\tau,n)$ integrates EEG over a 500 ms window. This ensures that we capture important temporal information in the EEG that relates to each sample of the stimulus we are trying to reconstruct. To quantify the contribution of each time lag towards reconstruction, decoders were trained on EEG at individual lags from 0–500 ms, instead of integrating across them (O'Sullivan et al., 2015). For a sampling frequency of 64 Hz, this equates to 33 individual lags and thus 33 separate decoders. For each time lag, the solution then becomes:

$$\hat{s}(t) = \sum_{n=1}^{128} r(t+\tau,n)g(\tau,n), \quad 0 < \tau \leq 500\mathrm{ms} \tag{5.3}$$

where $\hat{s}_{\tau}(t)$ is the estimated speech envelope for lag τ. Because the decoders consisted of only a single time lag, there was no need for regularization along the time dimension. Instead of using ridge regression to compute the decoder, it was approximated by performing a singular value decomposition of the auto-correlation matrix (Mesgarani et al., 2009, Ding and Simon, 2012b, Theunissen et al., 2000). Here, only those eigenvalues that exceed a specific fraction of the largest eigenvalue or peak power are included in the analysis. Qualitatively, this approach yields the same result as doing ridge regression but. To examine how MSI$_{\mathrm{EEG}}$ varied as a function of time lag, it was calculated as before (Eq. 5.2) using the single-lag decoders. To investigate whether MSI$_{\mathrm{EEG}}$ was predictive of MSI$_{\mathrm{Behav}}$ at a particular time lag, we calculated the correlation coefficient between the two measures across participants. This was examined in speech-in-noise, where behavioural performance was not at ceiling.

5.2.8 Statistical Analyses

All statistical analyses were conducted using two-way mixed ANOVAs with a between-subjects factor of SNR (Quiet vs −9 dB) and a within-subjects factor of condition (A, V, A+V, AV), except where otherwise stated. Where sphericity was violated in factors with two or more levels, the Greenhouse-Geisser corrected degrees of freedom are reported. *Post hoc* comparisons were conducted using two-tailed *t*-tests and multiple comparisons were corrected for using the Holm-Bonferroni method. All numerical values are reported as mean ± SD. Outlying participants were excluded from specific analyses if their values within that analysis were a distance of more than three times the interquartile range.

5.3 Results

5.3.1 Behaviour

Subjectively-rated intelligibility in the speech-in-noise experiment confirmed that intelligibility was highest in the AV condition ($t_{(20)} = 10.3$, $p = 1.9\times10^{-9}$; A+V: 36.9 ± 18.4%; AV: 63.6 ± 15.8%, Fig. 5.2*B*). This was reflected in how accurately participants could detect the target words, with detection accuracy significantly higher in the AV condition compared to that predicted by the unisensory scores ($t_{(20)} = 2.6$, $p = 0.018$; $\hat{F}_1(\mathrm{AV})$: 0.7 ± 0.09, $F_1(\mathrm{AV})$: 0.76 ± 0.08; Fig. 5.2*C*, left). In speech-in-quiet, accuracy in the A and AV conditions was at ceiling, hence there was no observable multisensory benefit. As a result, the AV gain for speech-in-noise was significantly greater than that for speech-in-quiet [unpaired *t*-test: $t_{(39)} = 2.8$, $p = 0.0086$; $\mathrm{MSI}_{\mathrm{Behav}}$ (Quiet): −1.44 ± 5.61%, $\mathrm{MSI}_{\mathrm{Behav}}$ (−9 dB): 9.14 ± 15.12%; Fig. 5.2*C*, right]. For speech-in-noise, both intelligibility and detection accuracy varied substantially across participants. Importantly, the individual accuracy scores were shown to be significantly correlated with intelligibility in both the unisensory conditions (A: $r = 0.51$, $p = 0.02$; V: $r = 0.55$, $p = 0.01$). In the AV condition, accuracy rates were nearer to ceiling, thus, any observable correlation with intelligibility was most likely obscured.

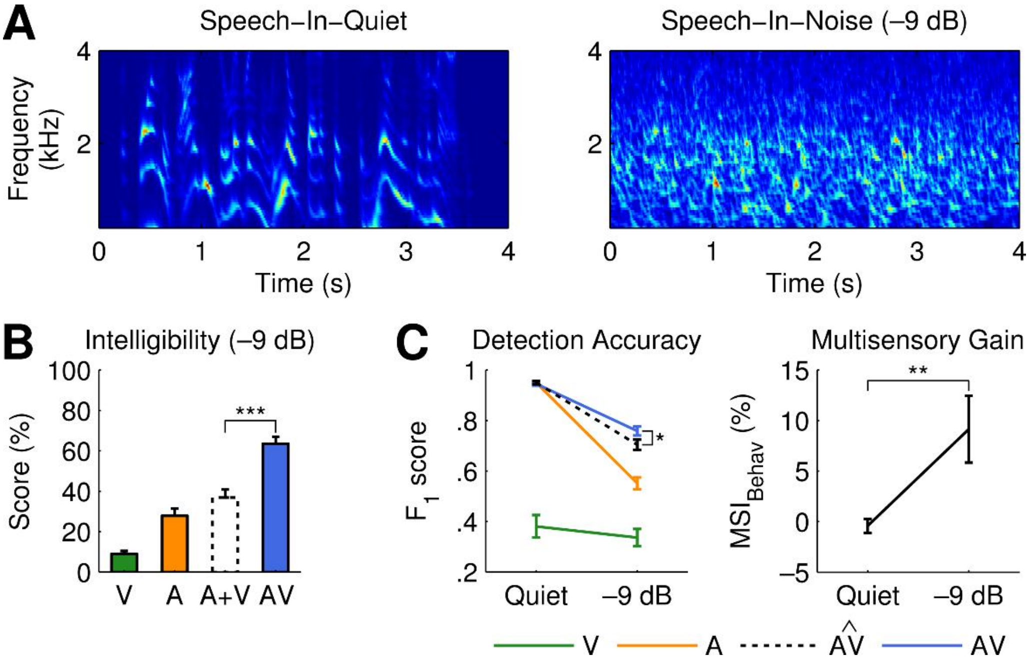


Figure 5.2: Audio stimuli and behavioural measures.

A, Spectrograms of a 4-s segment of speech-in-quiet (left) and speech-in-noise (−9 dB; right). ***B***, Subjectively-rated intelligibility for speech-in-noise, reported after each 60-s trial. The white bar represents the sum of the unisensory scores. Error bars indicate SEM across participants. Brackets indicate pairwise statistical comparisons (*$p < 0.05$;**$p < 0.01$; ***$p < 0.001$). ***C***, Detection accuracy (left) of target words represented as F1 scores. The dashed black trace represents the statistical facilitation predicted by the unisensory scores. Multisensory gain (right) represented as a percentage of unisensory performance.

5.3.2 Neural Enhancement and Inverse Effectiveness

Neural tracking of the speech signal was measured based on how accurately the broadband envelope could be reconstructed from the participants' EEG (Fig. 5.3*A*, left). A mixed ANOVA with factors of SNR (Quiet vs −9 dB) and condition (A vs V) revealed a significant interaction effect ($F_{(1,40)} = 24.1$, $p = 1.6\times10^{-5}$), driven by the fact that reconstruction accuracy in the A condition fell below that of the V condition at −9 dB SNR ($t_{(20)} = 2$, $p = 0.055$; A: 0.17 ± 0.05, V: 0.13 ± 0.04). Multisensory integration was indexed by differences in reconstruction accuracy between the AV condition and the A+V model. There was a main effect of condition across SNRs ($F_{(1,40)} = 115.1$, $p = 2.4\times10^{-13}$), with significantly higher reconstruction accuracy in the AV condition for both speech-in-quiet ($t_{(20)} = 7.1$, $p = 7.3\times10^{-7}$; AV: 0.2 ± 0.04, A+V: 0.18 ± 0.04) and speech-in-noise ($t_{(20)}$ = 8.1, p = 1×10^{-7}; AV: 0.16 ± 0.05, A+V: 0.14 ± 0.05). Although there was no significant interaction between SNR and condition ($F_{(1,40)}$ = 2.5, p = 0.12), the multisensory gain (i.e., the AV enhancement as a percentage of A+V) was significantly greater at −9 dB SNR than in quiet [unpaired t-test: $t_{(20)}$ = 2.8, p = 0.008; MSI$_{\text{EEG}}$ (Quiet): 10.6 ± 6.8%, MSI$_{\text{EEG}}$ (−9 dB): 20.7 ± 14.9%; Fig. 5.3*A*, right]. These findings demonstrate that envelope tracking is restored in adverse hearing conditions by the addition of visual speech and that this process conforms to the principle of inverse effectiveness.

To examine the time lags that contributed most towards reconstruction, 33 separate estimates of the speech envelope were reconstructed using single-lag decoders between 0–500 ms (Fig. 5.3*B*). The time lags that contributed the most information peaked earlier for speech-in-quiet (~110 ms) than for speech-in-noise (~170 ms). A running t-test comparing AV with A+V at each time lag indicated that multisensory interactions occurred at multiple stages (p < 0.05, Holm-Bonferroni corrected) and onset earlier for speech-in-quiet (~45 ms) than for speech-in-noise (~110 ms). For speech-in-noise, reconstruction accuracy in the A condition was significantly lower than that of the V condition between 0–95 ms (running t-test: p < 0.05, Holm-Bonferroni corrected). This suggests that in adverse hearing conditions, the sensitivity of auditory cortex to speech is significantly reduced during an early stage of speech processing.

5.3.3 Neural Enhancement Predicts Behavioural Gain

In order to investigate the relationship between our neural and behavioural measure of multisensory integration, we calculated the correlation coefficient between them using the reconstructed estimates from each of the 33 single-lag decoders. The logic here was that our behavioural multisensory effect may be reflected in our neural measure at a specific latency and integrating across 500 ms may obscure any correlation between these measures. Figure 5.3*C* shows the correlation between MSI$_{\text{Behav}}$ and MSI$_{\text{EEG}}$ at every time lag between 0–500 ms. There is no meaningful correlation for the first 200 ms, after which it begins to steadily increase until it peaks between 220–250 ms, at which latencies there is a significant (and positive) correlation (r = 0.44, p = 0.04; Fig. 5.3*D*, left). This correlation is also significant if MSI is represented as percentage gain (r = 0.56, p = 0.009; Fig. 5.3*D*, right). If we calculate a linear fit to these data, the slope of the resulting line is approximately 0.96, meaning that on average, a 50% gain in envelope tracking reflects a 52% gain in detection accuracy.

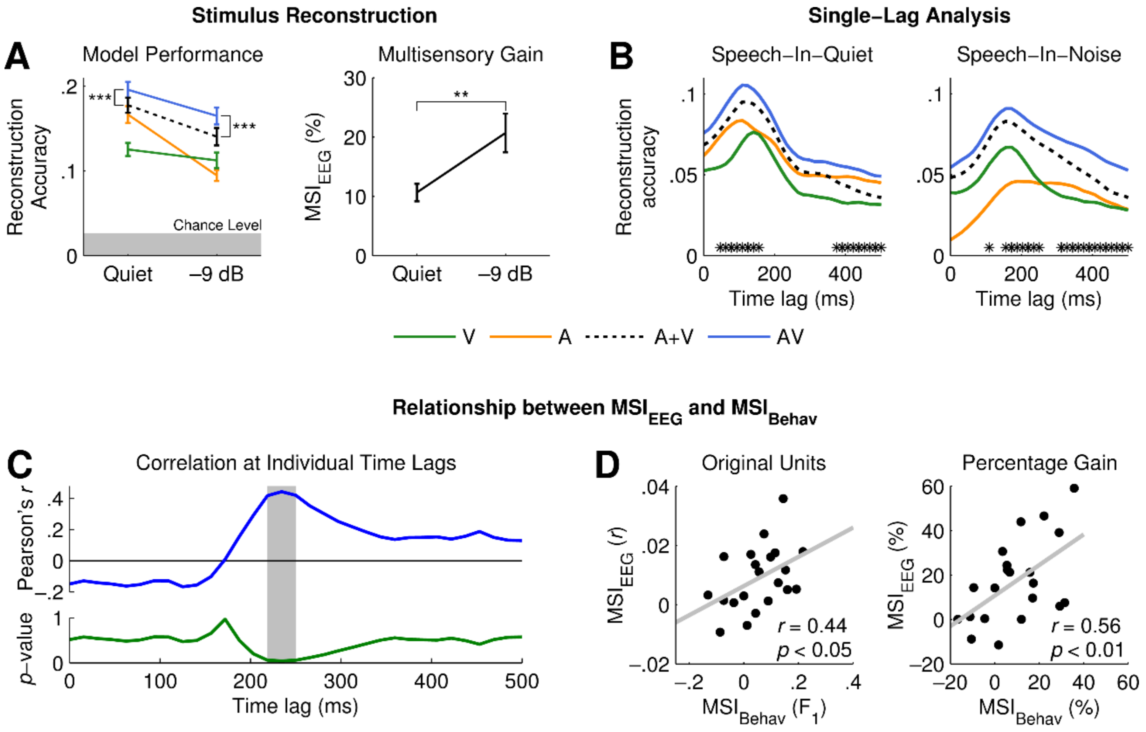


Figure 5.3: Stimulus reconstruction and relationship with behaviour.

A, Reconstruction accuracy (left) obtained using decoders that integrated EEG across a 500-ms window. The dashed black trace represents the unisensory additive model. The shaded area indicates the 95th percentile of chance-level reconstruction accuracy (permutation test). Multisensory gain (right) represented as a percentage of unisensory performance. Error bars indicate SEM across participants. Brackets indicate pairwise statistical comparisons (**$p < 0.01$; ***$p < 0.001$). ***B***, Reconstruction accuracy obtained using single-lag decoders at every lag between 0 and 500 ms. The markers running along the bottom of each plot indicate the time lags at which $\mathrm{MSI}_{\mathrm{EEG}}$ is significant ($p <$ 0.05, Holm-Bonferroni corrected). ***C***, Correlation coefficient (top) and corresponding *p*-value (bottom) between $\mathrm{MSI}_{\mathrm{EEG}}$ and $\mathrm{MSI}_{\mathrm{Behav}}$ at individual time lags for speech-in-noise. The shaded area indicates the lags at which the correlation is significant or trending towards significance (220–250 ms; $p < 0.05$). ***D***, Correlation corresponding to shaded area in ***C***, with $\mathrm{MSI}_{\mathrm{EEG}}$ and $\mathrm{MSI}_{\mathrm{Behav}}$ represented in their original units (left) and as percentage gain (right).

5.3.4 AV Speech Integration at Multiple Timescales

As demonstrated in Chapter 4, the timescale of AV speech processing is closely linked to the rate at which syllables occur in extended passages of natural speech (Chandrasekaran et al., 2009, Luo et al., 2010). To examine the impact of background noise on the timescale at which AV speech is integrated, we calculated the correlation coefficient between the reconstructed and original envelope at every 1-Hz frequency band between 1–30 Hz. Figure 5.4*A* shows the spectral profile of reconstruction accuracy for the AV condition and the A+V model. This spectrum represents the contribution of each frequency band to reconstructing the broadband envelope. Because the spectrum is consistently low-pass in shape, we defined the cutoff frequency as the highest frequency at which reconstruction accuracy was greater than chance level (permutation test). For speech-in-quiet, reconstruction accuracy was greater than chance at frequencies between 1–8 Hz (Fig. 5.4*A*, left), whereas for speech-in-noise, reconstruction accuracy was only greater than chance between 1–5 Hz (Fig. 5.4*A*, right).

Figure 5.4*B* shows the multisensory enhancement measured at each frequency band. To test for significance, paired *t*-tests were conducted at only the frequencies where reconstruction accuracy was greater than chance level ($p < 0.05$, Holm-Bonferroni corrected). For speech-in-quiet, there was a significant AV enhancement between 1–6 Hz (Fig. 5.4*B*, top), whereas for speech-in-noise, there was only a significant enhancement between 1–3 Hz (Fig. 5.4*B*, bottom). To relate these findings to the temporal scale of natural speech, we summarized the average rate of different linguistic units by deriving the durations of the respective speech segments from the audio files (Fig. 5.4*C*). The results suggest that in quiet, AV speech was integrated at frequencies commensurate with the rate of suprasegmental information such as sentential and phrasal units, as well as smaller segmental units such as words and syllables. In background noise, AV integration was only evident at the sentential and lexical timescale.

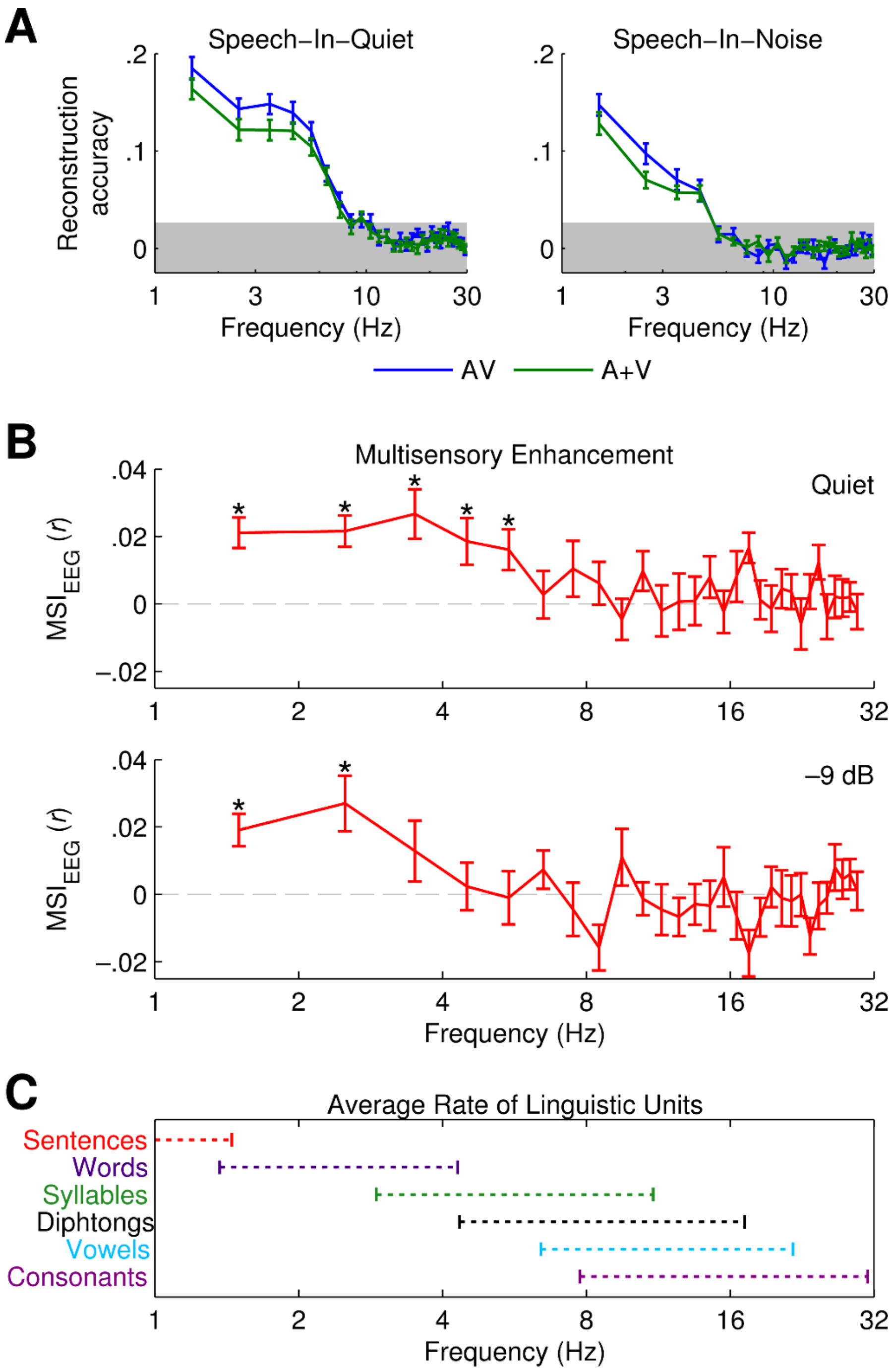


Figure 5.4: AV speech integration at multiple timescales.

A, Reconstruction accuracy for AV (blue) and A+V (green) at each frequency band. The shaded area indicates the 5th to 95th percentile of chance-level reconstruction accuracy (permutation test). Error bars indicate SEM across participants. ***B***, Multisensory enhancement at each frequency band. The markers indicate frequency bands at which there was a significant multisensory interaction effect (p < 0.05, Holm-Bonferroni corrected). ***C***, Average rate of different linguistic units derived from the audio files of the speech stimuli. The brackets indicate the mean ± SD.

5.3.5 Long-Term AV Temporal Integration

Given that background-insensitive speech recognition has been linked to long-term temporal integration (Ding and Simon, 2013), we wished to examine the role of temporal integration in maintaining AV speech processing in background noise. The decoder window size was shortened from 500 to 100 ms in steps of 100 ms, restricting the amount of temporal information that each decoder could integrate across when reconstructing the stimulus. While this reduced decoder performance in both quiet (ΔAV: 0.04 ± 0.01) and in noise (ΔAV: 0.06 ± 0.03), the effect was significantly greater in the latter (unpaired *t*-test: $t_{(40)} = 2.7$, $p = 0.01$; Fig 5.5*A*). As a result, multisensory gain was more sensitive to modulations in temporal window size in noise ($F_{(1.8,36.5)} = 1.4$, $p = 0.27$, one-way ANOVA) than in quiet ($F_{(1.3,26.7)} = 0.31$, $p = 0.87$, one-way ANOVA). Although the effect was not significant, $\mathrm{MSI_{EEG}}$ decreased as the temporal window size was reduced (see Fig. 5.5*B*). Critically, inverse effectiveness (i.e., the difference between $\mathrm{MSI_{EEG}}$ in quiet and noise) was only significantly greater than zero when the decoders integrated EEG over temporal window sizes of >300 ms (unpaired *t*-tests: $p < 0.05$; Fig. 5.5*C*).

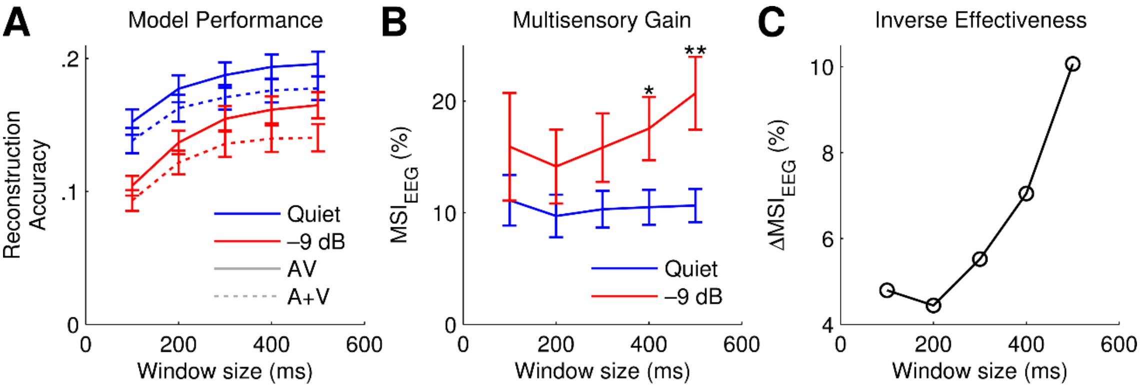


Figure 5.5: AV speech integration by temporal window size.

A, Model performance by decoder temporal window size. Error bars indicate SEM across participants. ***B***, Multisensory gain by decoder temporal window size. The markers indicate window sizes at which there was significant inverse effectiveness (i.e., −9 dB > Quiet; **p* < 0.05; ***p* < 0.01). ***C***, Inverse effectiveness by decoder temporal window size.

5.3.6 Left-Dominant Multisensory Interactions

The stimulus reconstruction approach utilises all 128 channels of EEG in order to maximise the variance captured across the scalp. To investigate whether or not our multisensory effect was lateralised, we repeated the analysis using channels from only the left and right sides of the head separately. Note that the decoders were not trained on the left and right channels separately; instead, they were trained on all 128 channels together, and during the reconstruction phase they were limited to the 53 left-most and 53 right-most channels (see Fig 5.6*C*). Although reconstruction accuracy was reduced by training on more channels than were used for reconstruction, the advantage of this approach is that the decoder does not allocate resources to the encoding of correlated input features from contralateral channels and thus will more likely reveal any lateralized effects that may be present. Additionally, the decoders were limited to time lags between 0–300 ms, which were shown to contribute most information in our single-lag analysis (Fig. 5.6*B*).

A three-way mixed ANOVA with a between-subjects factor of SNR and within-subjects factors of condition and hemisphere found no main effect of hemisphere ($F_{(1,39)} = 0.01$, $p = 0.91$; Fig 5.6*A*) or any interaction between SNR and hemisphere ($F_{(1,39)} = 0.4$, $p = 0.53$). Examining multisensory enhancement, a two-way mixed ANOVA (SNR × hemisphere) indicated a trend towards a main effect of hemisphere ($F_{(1,39)} = 3.7$, $p = 0.06$; Fig 5.6*B*) such that there was a marginally greater multisensory enhancement in left hemisphere. However, there was a significant interaction between SNR and hemisphere ($F_{(1,39)} = 5.4$, $p = 0.025$), driven by greater multisensory enhancement in the left hemisphere for speech-in-noise ($t_{(20)} = 3.3$, $p = 0.004$).

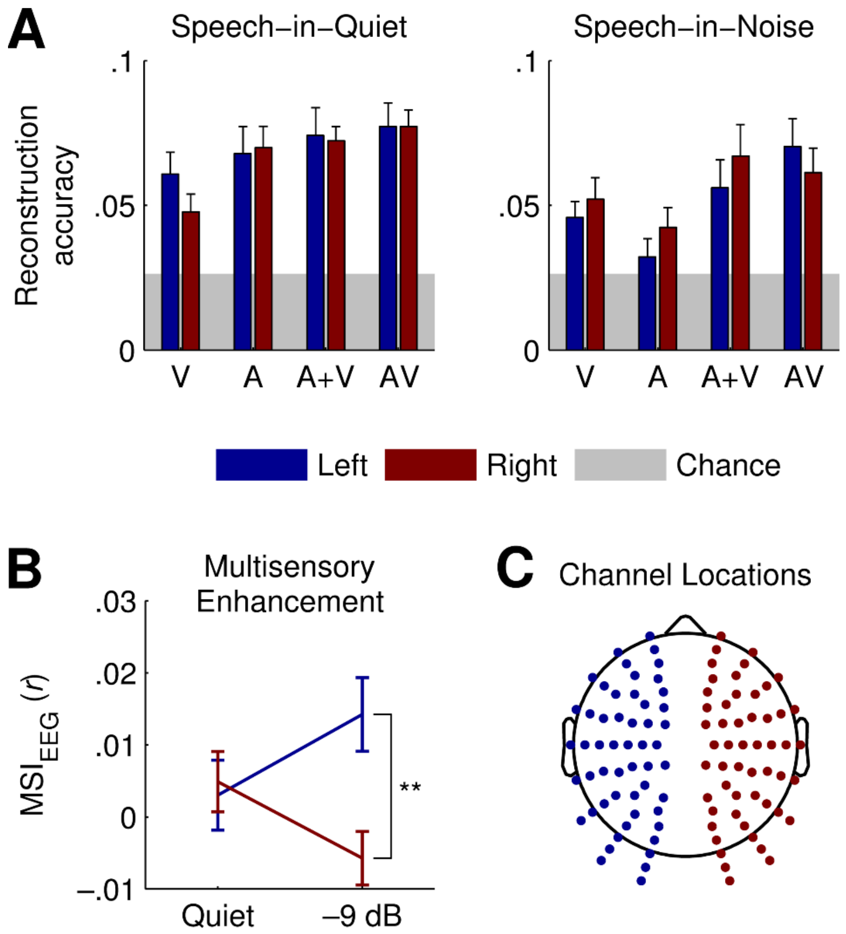


Figure 5.6: Multisensory enhancement by left and right hemisphere.

A, Reconstruction accuracy using the left (blue) and right (red) channels. The shaded area indicates the 95th percentile of chance-level reconstruction accuracy (permutation test). Error bars indicate SEM across participants. ***B***, Multisensory enhancement by hemisphere and condition. Brackets indicate pairwise statistical comparisons (***p* < 0.01). ***C***, Scalp locations of the electrodes chosen to represent the left and right hemispheres. Electrodes along the median line were excluded from the analysis.

5.4 Discussion

Our findings exhibit three major electrophysiological features of AV speech processing. First, the accuracy with which cortical activity entrains to AV speech conforms to the principle of inverse effectiveness. Second, visual speech input restores early tracking of the audio speech signal in noise and is integrated with auditory information at much lower frequencies. Third, inverse effectiveness in natural AV speech processing relies on crossmodal integration over long (>300 ms) temporal windows. Our findings suggest that AV speech integration is maintained in background noise by several underlying mechanisms, occurring at different temporal stages.

5.4.1 Quantifying a behavioural measure of MSI

From our measure of behavioural performance that was shown to reflect intelligibility, we sought to obtain an index of multisensory integration. However, isolating contributions from multisensory interactions can be obscured by artificially high speech-reading scores (Bernstein et al., 2004a, Ross et al., 2007a). Here, we attempted to circumvent this by accounting for false alarms and the likelihood that a target was detected in both modalities (Stevenson et al., 2014a). Using this probabilistic model, we were able to demonstrate that recognition accuracy in background noise was enhanced beyond the statistical facilitation predicted by independent unisensory processing (Blamey et al., 1989). In contrast, studies that have predicted AV performance based on “optimal processing” models typically yield predictions greater than participants’ observed AV performance (Grant et al., 2007, Braida, 1991).

Characterizing multisensory enhancement in terms of perceptual gain can also be achieved using a variety of methods (Ross et al., 2007a). Here, we chose to calculate gain as a percentage of unisensory performance. In doing so, we demonstrated that the gain was significantly greater in noise than it was in quiet, in line with Ross et al. (2007a). However, in their study they quantified gain as a percentage of auditory-only performance and remark that this approach is constrained by a ceiling effect at lower SNRs due to the inverse relationship between gain and auditory-only performance (Grant and Walden, 1996). Our approach, which also accounts for visual-only performance, is less prone to such ceiling effects.

5.4.2 Envelope Tracking and Inverse Effectiveness

In line with seminal work on AV speech-in-noise (Ross et al., 2007a, Sumby and Pollack, 1954), we demonstrated that the behavioural benefit produced by AV speech was significantly greater in noise than in quiet. This inverse effectiveness phenomenon was also observed in our EEG data, which revealed that multisensory interactions were contributing to the neural tracking of AV speech to a greater extent in noise than in quiet. In support of our neuronal effect, a recent MEG study demonstrated (using a phase-based measure of neural tracking) that coherence across multiple neural response trials was enhanced by AV speech relative to A speech when participants listened to competing-speakers, but not single-speakers (Zion-Golumbic et al., 2013a). In other words, making it more difficult to hear the target speaker by introducing a second speaker revealed an enhancement in AV speech tracking that was not detectable in single-speaker speech.

For speech-in-noise, we found that the multisensory enhancement in envelope tracking at 220–250 ms accurately predicted the multisensory gain in behaviour. To interpret the significance of this temporal locus, we must first consider what these multisensory indices reflect. Our behavioural measure ($\mathrm{MSI}_{\mathrm{Behav}}$) was derived from the accuracy with which participants detected target words. Because the task involved identifying whole words, the MSI score may reflect integration at the semantic level (Ross et al., 2007a). In support of this, the time course of speech perception in the superior temporal cortex has been shown to reflect lexical-semantic processing from 200 ms onwards (Salmelin, 2007, Picton, 2013). Our neural measure ($\mathrm{MSI}_{\mathrm{EEG}}$), on the other hand, was derived from how accurately the speech envelope could be reconstructed from the EEG data. Specifically, we observed multisensory interactions below 3 Hz in noise. Given that this frequency range is commensurate with the average rate of spoken words, it fits well with our behavioural task. Furthermore, neural oscillations in the delta range (1–4 Hz) are thought to integrate crossmodal information over a temporal window of ~125–250 ms (Schroeder et al., 2008), in line with the latency of our effect. Together, this suggests that our neural and behavioural measure of multisensory integration both reflect processing at the lexical-semantic level and, as such, are correlated at a timescale that corresponds to this stage of speech processing.

5.4.3 Neural Mechanisms in AV Speech-in-Noise

Our EEG data suggest that cortical activity entrains to AV speech only at lower frequencies in background noise. In support of this notion, it has been demonstrated that MEG entrains to AV speech at lower frequencies when a competing speaker is introduced (Zion-Golumbic et al., 2013a). An MEG study by Ding and Simon (2013) that investigated neural entrainment to audio-only speech at different SNRs found that the cutoff frequency of the phase-locking spectrum decreased linearly with SNR, but that low delta-band neural entrainment was relatively insensitive to background noise above a certain threshold. This mechanism of contrast gain control was linked to the

M100 component of the temporal response function (TRF), which was shown to be relatively robust to noise, unlike the earlier M50 component (Ding and Simon, 2013, Ding et al., 2014). Our results, along with these other studies, indicate that low-frequency speech information is more reliably encoded than higher-frequency linguistic content in adverse hearing conditions and that this process is maintained by contrast gain control and adaptive temporal sensitivity in auditory cortex (Ding and Simon, 2013).

In addition, we found that auditory and visual information interacted at lower frequencies in noise than in quiet, which is unsurprising, given that there is a more robust auditory representation encoded at lower frequencies. In line with this, we showed that inverse effectiveness relied on longer temporal windows of integration, something that is also critical for a noise-robust cortical representation of speech (Ding and Simon, 2013). A recent intracranial study that examined AV integration in quiet using discrete, non-speech stimuli, observed multisensory enhancement effects [AV−(A+V)] in delta and theta phase alignment (Mercier et al., 2015). Interestingly, they reported visually driven crossmodal delta-band phase-reset in auditory cortex. It is possible that this process could be mediated by delta-frequency head movements, which have been shown to convey prosodic information important to speech intelligibility (Munhall et al., 2004b). Thus, integration of auditory and visual speech information could be maintained in adverse hearing conditions by a combination of delta-frequency phase-resetting and long-term temporal integration.

Chapter 6 Investigating the Temporal Dynamics of Auditory Cortical Activation to Silent Lipreading

6.1 Introduction

Functional neuroimaging research has demonstrated that observing visual speech (i.e., lipreading) in the absence of auditory speech activates primary auditory cortex in humans (Calvert et al., 1997, Pekkola et al., 2005). It has also been shown using single-unit recordings in the primate brain that visually-presented monkey articulations can elicit local field potentials in auditory cortex (Kayser et al., 2008). Indeed, fMRI has greatly advanced our understanding of multisensory integration in the human brain and, in particular, where in the brain it occurs. However, because of its poor temporal resolution, fMRI is not well suited to examining the neural response to dynamic speech stimuli that rapidly fluctuate over time. Thus, it is difficult to determine what this auditory cortical activation during silent lipreading precisely reflects. While single-unit recordings in primates offer better temporal resolution, monkey vocalisations are not directly comparable to human speech, as they lack the lexical complexity of human conversation. Essentially, studying the dynamics of visual speech processing is better suited to non-invasive human recording techniques with high temporal resolution such as EEG and MEG.

As discussed previously, such techniques have been instrumental in demonstrating that auditory cortical activity tracks the temporal envelope of acoustic speech (Lalor and Foxe, 2010, Ding and Simon, 2012b). However, Ding et al. (2014) suggest that envelope tracking may actually reflect an analysis-by-synthesis process, whereby speech features that are correlated with the envelope are encoded during the synthesis phase, thus leading to what appears to be merely tracking of the speech envelope (see section 2.4.4). Given that many of the visual cues that contribute to speechreading are also correlated with the acoustic envelope (Chandrasekaran et al., 2009, Grant and Seitz, 2000), encoding of such features could theoretically manifest in a process that also time-locks to the speech envelope. If the temporal dynamics of such visual cues are projected to auditory cortical regions during silent lipreading, this could elicit envelope tracking in auditory cortex in the absence of acoustic speech. Here, this hypothesis is tested by examining the impact of lipreading accuracy on the entrainment of EEG to the unheard speech signal.

In Chapters 3, 4 and 5, we demonstrated that it is possible to reconstruct an estimate of the speech envelope from EEG data. While this work has distinct applications in brain-computer interface technology, such methods would better serve BCIs by decoding the users' inner thoughts, i.e., covert speech. Such an approach presents two main challenges: (1) how do we model the neural representation of an internal process, and (2) how do we determine the exact time at which it occurred? Recently, Martin et al. (2014) successfully decoded covert speech from ECoG data using a decoder that modelled the neural representation of overt speech, while timing issues were dealt with using dynamic time warping. The present chapter demonstrates how the natural statistics of visual speech can be utilised to overcome these issues: (1) assuming that speech perception and imagery share a partially overlapping cortical representation, the original acoustic signal can be used as an estimate of what the perceiver imagined, and (2) timing issues are naturally circumvented because the perceiver is continually prompted to imagine the auditory speech content, time-locked to the visual cues. Here, both forward and backward modelling techniques are applied as quantitative measures of envelope tracking during silent lipreading. The results of this study were presented at the *5th International Conference on Auditory Cortex* in Magdeburg in September, 2014 (Appendix J) and published in *Proceedings of the 7th International IEEE/EMBS Conference on Neural Engineering* (Crosse et al., 2015b).

6.2 Methods

6.2.1 Participants

Twelve native English speakers (5 females; age range: 22–37 years), none of which were trained lipreaders, gave written informed consent. All participants were right-handed, free of neurological diseases, had normal hearing and normal or corrected-to-normal vision. The experiment was undertaken in accordance with the Declaration of Helsinki and was approved by the Ethics Committee of the Health Sciences Faculty at Trinity College Dublin, Ireland.

6.2.2 Stimuli and Procedure

The speech stimuli used for the lipreading experiment were the same videos used in the visual-only (V) and congruent audiovisual (AV) conditions in Chapter 4. Soundtracks were removed from 14 of the 15 videos used for silent lipreading. The remaining video was preserved in AV format and used as a control. Please refer to Chapter 4 for exact video specifications (see section 4.2.2).

Task instructions and testing conditions were identical to those described in section 4.2.2. Prior to EEG testing, participants were trained on the AV stimulus to ensure familiarity with the speech content. During EEG testing, the same AV stimulus was presented 14 times as a control. This known video (Vk) was also presented in visual-only format 14 times, for which participants were instructed to lipread. The remaining 14 unknown videos (Vu) were presented once each in visual-only format. Participants were instructed to lipread the Vu stimuli even though they were not familiar with the audio content. Stimulus presentation order was randomised across conditions, within participants. During each 60-s trial, participants were required to respond to a target word with a button press. A different set of target words was used for each condition and the assignment of target words was counterbalanced across participants. Each target word occurred between 1 and 3 times per trial and there were 28 targets in total per condition.

6.2.3 EEG Acquisition and Pre-Processing

EEG data were acquired using the same high-density recording system and experimental protocol described in section 4.2.4. Pre-processing conducted offline, including filtering and artefact rejection, was identical to that described in section 4.2.4.

6.2.4 Stimulus Reconstruction

To obtain a quantitative measure of envelope tracking, stimulus reconstruction was conducted as described in Chapter 3 (see section 3.2.4). Decoders were optimised using the same specifications described in Chapter 4 (see section 4.2.6). The objective, still, was to reconstruct the acoustic speech envelope, even though the participants did not hear it whilst lipreading. The rational here is that the temporal dynamics of the absent acoustic signal might still be reflected in auditory cortical areas due to correlated phasic variations projected from visual cortex (Luo et al., 2010).

For each subject, a separate search of the ridge parameter was conducted (2^{10}, 2^{11},…, 2^{30}) such that reconstruction accuracy was optimised within each condition. The ridge value with the highest mean reconstruction accuracy over the 14 trials was chosen to prevent overfitting (see section 3.3.2). However, in the AV and Vk conditions, the same stimulus was repeated over the 14 trials, which may have caused overfitting. An additional analysis was included which removed any potential bias by using 'grand-average' decoders as opposed to 'subject-specific' decoders (Crosse et al., 2013, O'Sullivan et al., 2015). Essentially, instead of conducting a leave-one-out cross-validation on the 14 trials for each subject, it was conducted across participants for each particular trial. While unbiased, this approach yields a more generalised model and hence does not perform as well as a subject-specific model.

6.2.5 Temporal Response Function Estimation

To examine the relationship between the neural response and the presented stimulus, the temporal response function for each of the three conditions was calculated as described in Chapters 3 and 4 (see sections 3.2.1 and 4.2.8). Because the TRF mappings between the envelope and EEG were estimated in response to visual speech, there was no need to model the auditory periphery. Thus, speech envelopes were extracted using a straight Hilbert transform as described in Chapter 3 (see section 3.4.1). The subsequent envelope estimate was filtered below 25 Hz and downsampled to 512 Hz.

6.2.6 Statistical Analyses

All statistical analyses were conducted using one-way repeated-measures ANOVAs and Greenhouse-Geisser correction was applied where necessary. *Post hoc* comparisons were made using two-tailed paired *t*-tests, except where otherwise stated. All numerical values are reported as mean ± SD, unless otherwise stated.

6.3 Results

6.3.1 Behaviour

Twelve participants performed a target detection task during EEG recording. RTs were measured from the onset of auditory voicing and hits were counted for responses that were made 200–2000 ms after target onset. Condition had a significant impact on both hit rate ($F_{(2,22)} = 76.2$, $p < 0.001$; Fig. 6.1*A*) and RT ($F_{(1.3,14.7)} = 24.2$, $p < 0.001$; Fig. 6.1*B*). Planned comparisons showed that participants were significantly more accurate in the Vk condition (74 ± 11%) compared to the Vu condition (33 ± 15%; $t_{(11)} = 9.2$, $p < 0.001$) and that RTs were faster for Vk (532 ± 123 ms) relative to Vu (787 ± 150 ms; $t_{(11)} = 7.0$, $p < 0.001$).

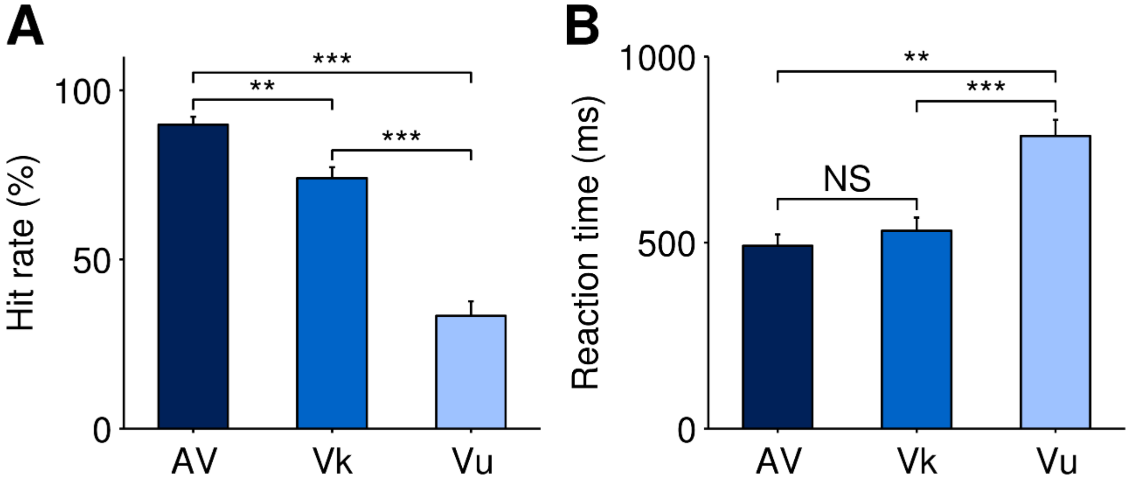


Figure 6.1: Behavioural performance.

A, Mean hit rates for AV, Vk and Vu speech. ***B***, Mean reaction times for all three conditions. Error bars indicate SEM across participants. Brackets indicate pairwise statistical comparisons (**p* < 0.05, ***p* < 0.01, ****p* < 0.001, NS = not significant).

6.3.2 Impact of Lipreading on Envelope Tracking

Stimulus reconstruction was applied using two different techniques. In the first, decoders were averaged across trials (within participants) as in Chapters 3 and 4. A one-way ANOVA revealed that condition had a significant impact on reconstruction accuracy ($F_{(2,22)} = 29.5$, $p < 0.001$; Fig. 6.2*A*). Critically, a *post hoc* comparison showed that reconstruction accuracy in the Vk condition (0.1 ± 0.03) was significantly higher than that of the Vu condition (0.08 ± 0.03; $t_{(11)} = 2.5$, $p < 0.05$). Although care was taken to optimise regularization within each condition, it remains a possibility that the conditions with repeated stimuli (AV and Vk) were somewhat biased. In the second analysis, this bias was removed by averaging the decoders across participants but within trials (Crosse et al., 2013, O'Sullivan et al., 2015). The main effect of condition on reconstruction accuracy was weakened by this approach ($F_{(2,22)} = 6.2$, $p < 0.01$; Fig. 6.2*B*). There was also no significant difference in reconstruction accuracy between the Vk condition (0.041 ± 0.02) and the Vu condition (0.044 ± 0.02; $t_{(11)} = 0.5$, $p > 0.05$). While mean reconstruction accuracy values were significantly reduced across all three conditions, they were still above chance level (see Fig. 6.2*B*).

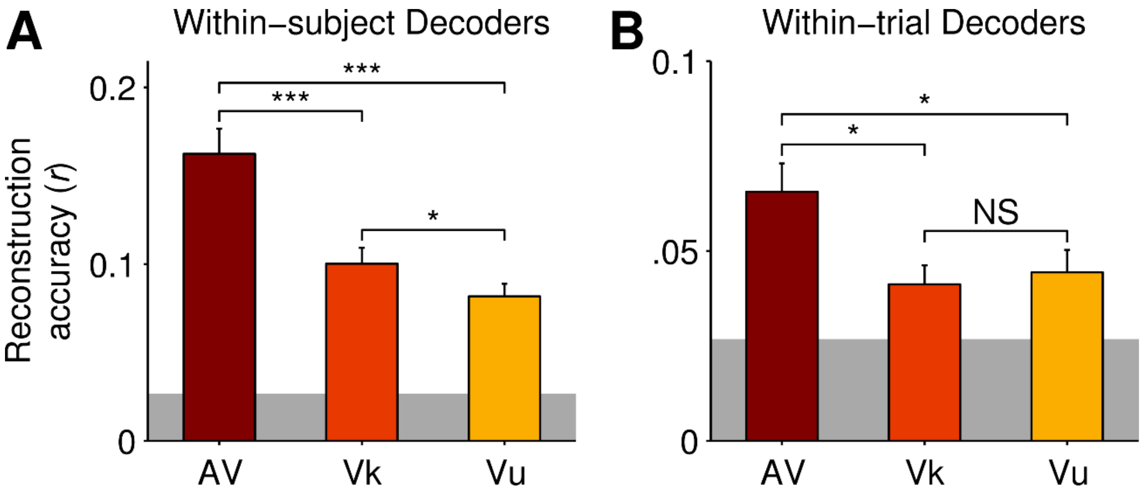


Figure 6.2: Reconstruction of the speech envelope during lipreading.

A, Mean reconstruction accuracy of decoders fit within participants, across trials. ***B***, Mean reconstruction accuracy of decoders fit within trials, across participants. The shaded area represents the 95th percentile of chance level (permutation test). Error bars indicate SEM across participants. Brackets indicate pairwise statistical comparisons (**p* < 0.05, ***p* < 0.01, ****p* < 0.001, NS = not significant).

6.3.3 Spatiotemporal Representation of Covert Speech

The EEG TRF (Lalor and Foxe, 2010) contains two major response components: a negativity at ~80 ms ($N1_{TRF}$) and a positivity at ~130 ms ($P2_{TRF}$; see Fig. 6.3*A*). Fig. 6.3*B* shows the topography of the $N1_{TRF}$ (left) and $P2_{TRF}$ (right) components. TRF SNR was defined as 0 to 250 ms (signal) and −100 to 0 ms (noise). Over fronto-temporal scalp, SNR was significantly lower in the Vk condition (1.3 ± 1 dB, mean ± SEM) and the Vu condition (0.45 ± 1 dB, mean ± SEM) compared to the AV condition (6.5 ± 1.9 dB, mean ± SEM; $F_{(2,22)} = 4.9$, $p < 0.05$; Fig. 6.3*A*, top) but was similar for all three over occipital scalp ($F_{(2,22)} = 0.07$, $p > 0.05$; Fig. 6.3*A*, bottom). This was reflected in the statistical cluster maps (Fig. 6.3*C*) which show significant activation across participants between 100–200 ms over parieto-occipital scalp in all three conditions and also over fronto-temporal scalp in the AV condition (running *t*-test: $p < 0.05$; Fig. 6.3*C*, top).

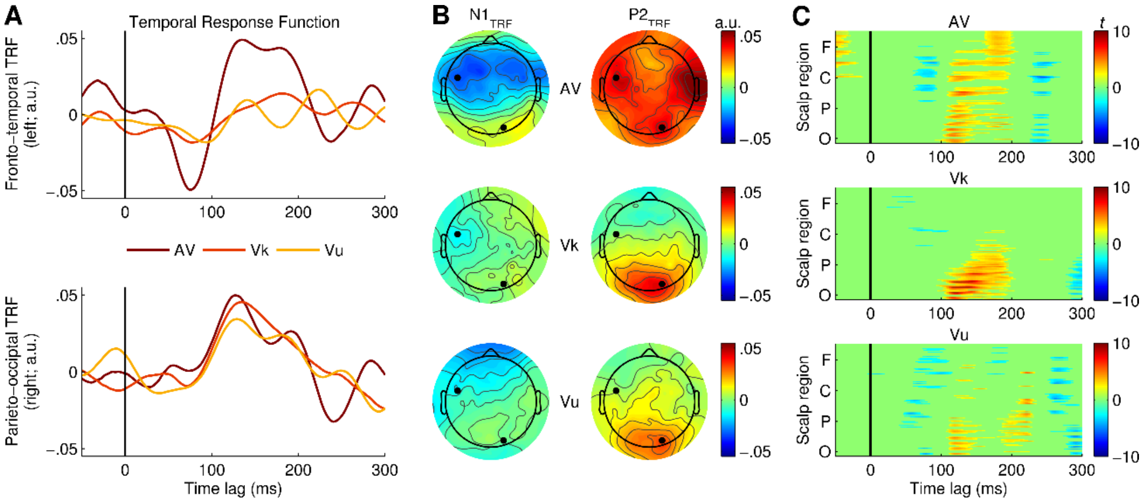


Figure 6.3: TRF timecourse and topography.

A, TRFs over left fronto-temporal scalp (top) and right parieto-occipital scalp (bottom). ***B***, Topographies of N1_{TRF} components (left) and P2_{TRF} components (right). Black markers indicate channel locations plotted in ***A***. ***C***, Statistical cluster maps show where and when TRF amplitude is significantly different to zero ($p < 0.05$; F = frontal, C = central, P = parietal, O = occipital).

6.3.4 Correspondence between TRF Dynamics

To compare the responses of the visual conditions (Vk and Vu) to those of the control condition (AV), a series of Pearson's correlations were performed on their TRFs between 0–250 ms. Fig. 6.4 shows the correlation coefficient (*r*) at each channel location. Channels where *r* is significantly greater than zero across participants are indicated by black markers (paired *t*-tests: $p < 0.05$). The comparison between Vk and Vu revealed a significant cluster of channels over occipital scalp (Fig. 6.4, right). Critically, there was also a cluster over left temporal scalp in the AV-Vk comparison (Fig. 6.4, left). The AV-Vu comparison, on the other hand, did not reveal any large clusters at channel locations where TRF activation was significant (Fig. 6.4, middle).

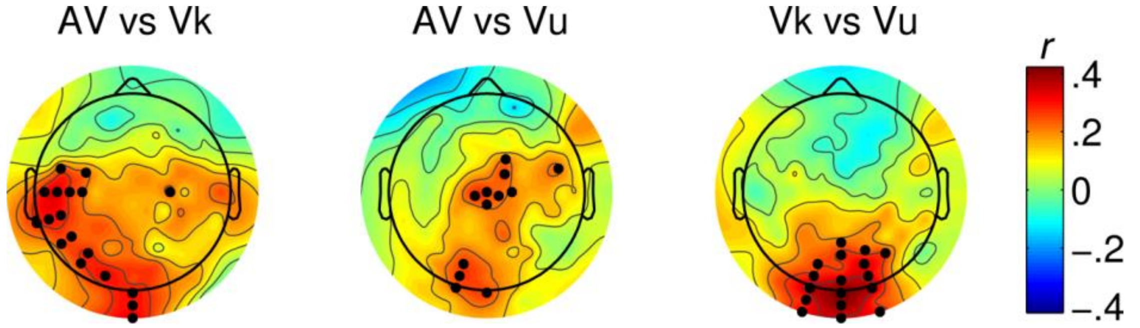


Figure 6.4: Comparison of TRF dynamics across conditions.

Mean correlation coefficients (*r*) between the TRFs (0–250 ms) of each condition at each channel location. Black markers indicate channels where *r* is significantly greater than zero across participants ($p < 0.05$).

6.4 Discussion

In this chapter, we tested the hypothesis that, during silent lipreading, auditory cortical activity synthesises the dynamics of the unheard speech signal by phase locking to inputs from the visual system. Specifically, we showed that envelope tracking during silent lipreading may be influenced by how accurately the participant could lipread. We also demonstrated that the temporal profile of the neural response to silent lipreading was significantly correlated with that of AV speech over left temporal scalp, but only when lipreading was accurately perceived.

6.4.1 Lipreading Accuracy and Envelope Tracking

The results of the within-subject decoding analysis indicate that successful lipreading results in better envelope tracking. This was true when each decoder was optimised separately within each condition so as not to bias those with repeated stimuli. However, as stated earlier, we cannot be certain that the AV and Vk decoders were not somewhat biased as a result of overfitting. Future work could address this issue by assigning a different stimulus to each condition (counterbalanced across participants) and presenting each one an equal number of times to ensure equal bias. In the absence of such experimental manipulations, a within-trial analysis was carried out which removed any potential bias from the AV and Vk conditions. However, because this approach involved averaging decoders across participants, the decoders were grossly generalised; thus, reconstruction accuracies dropped considerably and the effect was no longer significant. This is likely caused by the inherent spatiotemporal variability in the neural activity across the twelve participants (O'Sullivan et al., 2013). As such, it is difficult to definitively say whether there really was no difference between the Vk and Vu conditions, or whether the result was obscured by a floor effect. Furthermore, in support of the within-subject analysis, we demonstrated in Chapter 4 that envelope tracking was highly correlated with lipreading accuracy during visual-only speech (see Fig. 4.3*C*).

6.4.2 Speech-Reading in Auditory and Visual Cortical Regions

The TRF, which maps a continuous sensory input to the recorded neural response, was used as a direct measure of envelope tracking (Lalor and Foxe, 2010, Ding and Simon, 2012b). We found that, although TRF SNR was relatively low over fronto-temporal scalp during silent lipreading, its temporal profile was significantly correlated with that of AV speech when lipreading was accurately perceived. This may suggest that accurate processing of visual speech features plays a role in envelope tracking, in line with work espousing an analysis-by-synthesis framework (Ding et al., 2014). This is also supported by numerous studies that have reported attentional effects on envelope tracking (e.g., Power et al., 2012, Ding and Simon, 2012a, Mesgarani and Chang, 2012). Indeed, it is important to consider the possibility that using the same stimulus in two of the three conditions may have had an impact on the similarity of the TRFs. In theory, this should not influence the correlation between TRFs because a TRF represents the impulse response to a unit change in stimulus intensity (Lalor and Foxe, 2010). This is supported by the fact that the TRFs over occipital scalp were very similar in the Vk and Vu conditions (see Fig. 6.3, right), even though different visual stimuli were presented.

During silent lipreading, auditory cortex is not directly stimulated via the auditory nerve; hence TRF SNR over fronto-temporal scalp was reduced in the Vk and Vu conditions relative to the AV condition. This issue could be addressed in future work by presenting audio noise that is spectrally matched to the absent speech signal (Ding et al., 2014). This would directly stimulate the auditory nerve which may help boost auditory cortical responses. In other words, the visual cues could be used to ‘shape’ the acoustic noise into the speech signal that is being synthesised. There was no difference in TRF SNR over occipital scalp because each condition was matched in terms of visual stimulus intensity. The regression analysis is sensitive to this occipital activation because instantaneous measures of motion during visual speech are highly correlated with the amplitude of the acoustic envelope (Chandrasekaran et al., 2009, Grant and Seitz, 2000). However, in keeping with an analysis-by-synthesis model, this occipital activity may in fact reflect the processing of linguistic visual features in visual cortex as opposed to just motion tracking. It has been shown that every level of speech structure can be perceived visually, thus suggesting that there are visual modality-specific representations of speech in visual brain areas and not just in auditory brain areas (see Bernstein and Liebenthal, 2014).

Chapter 7 General Discussion

Within this body of research work, an intuitive and versatile SI toolbox has been developed for studying sensory processing in an electrophysiological context. A framework was established for studying multisensory integration in natural AV speech and was successfully implemented in three empirical studies. Moreover, this empirical work has yielded valuable insights into the neural basis of multisensory speech processing and, in particular, the adaptive spectrotemporal nature of AV speech integration.

Specifically, it was demonstrated in Chapter 4 that cortical activity entrains to speech more accurately when a listener can also see the speaker's face, despite the speech content being perfectly audible. Multisensory integration was observed at the syllabic timescale (2–6 Hz) and occurred at a late stage of speech processing (100–250 ms). It was also demonstrated that, during silent lipreading, entrainment to the dynamics of the absent acoustic signal (through the use of visual speech cues) was predictive of behavioural performance, suggesting that this process reflects higher-level speech processing and not merely motion tracking. Interestingly, envelope tracking was inhibited by AV stimuli that were incongruent both temporally and contextually. This effect is likely modulated by an overall reduction in crossmodal attention.

In Chapter 5, it was shown that neural tracking of the envelope during AV speech-in-noise conforms to the principle of inverse effectiveness. While envelope tracking during audio-only speech was greatly reduced by background noise at an early processing stage, it was markedly restored by the addition of visual speech input. In background noise, multisensory integration occurred at much lower frequencies and was shown to predict the multisensory gain in behavioural performance at a time lag of ~250 ms. Critically, we demonstrated that inverse effectiveness in natural audiovisual speech processing relies on crossmodal integration over long (>300 ms) temporal windows.

These results suggest that AV speech is integrated at a more coarse-grained (i.e., higher) linguistic level in adverse hearing conditions. Furthermore, neuronal multisensory interactions were predominantly left-lateralised for speech-in-noise. Given that speech processing becomes more left-lateralised further up the auditory hierarchy, this finding suggests that AV speech is integrated at a higher linguistic level in adverse listening environments. Together, my findings indicate that disparate integration mechanisms contribute to audiovisual speech processing in adverse hearing conditions over multiple stages.

In Chapter 6, EEG data recorded during silent lipreading suggest that more accurate speech-reading leads to improved global entrainment to the speech signal, in line with our findings in Chapter 4. Moreover, successful lipreading appeared to modulate cortical activity over left temporal scalp (i.e., near auditory cortex) and this activation was shown to reflect the dynamics of the absent acoustic speech envelope. We contend that silent lipreading may invoke phasic variations in auditory cortex reflective of the absent speech signal, akin to a synthesis of the covert auditory information. It is likely that such a process would aid cortical entrainment to the acoustic speech input in a multisensory context.

The rest of this chapter will focus on more general discussion that relates to the thesis as whole. Specifically, different theoretical frameworks of AV speech integration are discussed in the context of the thesis work, as well as the existing literature. The chapter concludes with discussion on the significance of this work and its future directions in scientific and clinical research.

7.1 Temporal Coherence as a Mechanism in AV Speech Processing

It has been suggested that the integration of auditory and visual speech may be driven by the temporal coherence of crossmodal information (Zion-Golumbic et al., 2013a). Computational and theoretical perspectives on stream segregation postulate that multi-feature auditory sources are segregated into perceptual objects based on the temporal coherence of the neuronal responses to the various acoustic features (Elhilali et al., 2009, Shamma et al., 2011, Ding and Simon, 2012a). Recently, Ding et al. (2014) demonstrated that cortical entrainment to the speech envelope does not reflect encoding of the envelope per se, as it relies on the spectrotemporal fine structure of speech. They suggest that it may instead index an analysis-by-synthesis mechanism, whereby spectrotemporal features that are correlated with the envelope are encoded during the synthesis phase (for a review, see Ding and Simon, 2014). In keeping with previous work espousing a correlated mode of AV speech (Campbell, 2008), the results presented in Chapter 4 indicate that visual speech cues, being correlated with the dynamics of the acoustic speech envelope, results in the visual signal being bound to the auditory features to form a multisensory object. In support of this notion, Rahne et al. (2008) demonstrated that two different tonal sequences separated by frequency were reliably perceived as one integrated stream when the accompanying visual stimulus was temporally coherent with both of them.

7.2 Brain Regions and Neural Mechanisms in AV Speech Processing

In terms of the specific neural populations that may facilitate the binding of temporally coherent visual and auditory speech, one candidate region is the superior temporal sulcus, which, as mentioned above, has previously been linked with multisensory object formation (see section 2.4.3; Beauchamp et al., 2004, Kayser and Logothetis, 2009, Calvert and Campbell, 2003). Indeed recent research has provided evidence for neural computations in this area that underpin auditory figure-ground segregation using stimuli that display periods of temporal coherence across multiple frequency channels (Teki et al., 2011). That the results presented here may derive from emergent activity during AV speech could suggest a role for the supramarginal and angular gyrus (Bernstein et al., 2008), although this particular study found these effects only in left hemisphere. Of course, in addition to such putatively multisensory regions, it remains a possibility that information pertaining to the timing of crossmodal stimuli could be projected to classic sensory-specific regions in a thalamocortical feedforward manner (Foxe and Schroeder, 2005, Besle et al., 2008) or laterally from other sensory-specific regions (Schroeder et al., 2008, Arnal et al., 2009, Besle et al., 2008). The long latencies of the multisensory effects presented in Chapter 4 may make this explanation less likely however, at least in the context of a correlated mode of AV speech.

A possible neural mechanism recently proposed also relates to the correlation between the speech envelope and visual motion. This theory suggests that anticipatory visual motion could produce phasic variations in visual cortical activity that are relayed to auditory cortex and that correlate with the amplitude envelope of the subsequent acoustic speech signal. This notion fits with MEG work which has demonstrated that the phase of oscillations in auditory cortex tracks the temporal structure of continuous visual speech (Luo et al., 2010) and fMRI work which has demonstrated that the source of the visual facilitation of auditory speech processing arises from motion-sensitive cortex (Arnal et al., 2009). Another suggestion for how visual speech may impact upon auditory speech processing is that this interaction may be driven by relatively discrete visual landmarks (e.g., the onset of facial articulatory movements) that elicit a phase-reset of ongoing low-frequency oscillations in auditory cortex, such that the arrival of the corresponding auditory syllable coincides with a high excitability phase of the auditory neuronal population (Kayser et al., 2008, Schroeder et al., 2008). The efficacy of such a mechanism in the context of continuous speech seems like it would necessitate prior knowledge about incoming information at the phonetic level. This process could in part be mediated by preceding visual cues which could elicit hierarchically organised phasic variations in visual cortex, continually updating auditory cortex prior to the arrival of the corresponding acoustic segment (see section 2.4.4).

7.3 An Analysis-By-Synthesis Perspective of Visual Speech Processing

In Chapter 4, it was demonstrated that it is possible to reconstruct an estimate of the acoustic envelope from visual speech data with accuracy well above chance level (Fig. 4.3*B*). Although the acoustic envelope was not explicitly encoded in the neural data during visual speech, it may still be inferred if some correlated feature of the visual speech was encoded (Mesgarani et al., 2009), as discussed above. One possible explanation is that instantaneous measures of motion during visual speech are highly correlated with the amplitude of the acoustic envelope (Chandrasekaran et al., 2009). However, in keeping with an analysis-by-synthesis framework, the data in Chapter 6 suggest that this occipital activity may in fact reflect the processing of higher-level (phoneme-level) visual speech features in visual cortex in addition to just motion tracking. It has been demonstrated that every level of speech structure can be perceived visually, thus suggesting that there are visual modality-specific representations of speech in visual brain areas and not just in auditory brain areas (for a review, see Bernstein and Liebenthal, 2014). Furthermore, a strong correlation was observed between behaviour and envelope tracking in the visual speech data (Fig. 4.3*C*), similar to that recently demonstrated in auditory speech-in-noise (Ding and Simon, 2013). As such, we tentatively suggest that lipreading accuracy is reflected in the neural tracking of the envelope, and that this tracking process includes the synthesis of visual speech tokens in visual-specific brain regions. While the challenges associated with using stimulus reconstruction to tease this issue apart have been outlined above, the use of different paradigms within our framework may yet prove enlightening.

7.4 Multistage Integration Model

As mentioned earlier, a growing body of evidence indicates that multisensory integration likely occurs over multiple temporal stages during AV speech processing (van Wassenhove et al., 2005, Eskelund et al., 2011, Baart et al., 2014, Schwartz et al., 2004, Peelle and Sommers, 2015). The findings presented in Chapters 4 and 5 and the findings of other studies will be interpreted within the context of such multistage integration models and, in particular, the role of prediction and constraint as early and late integration mechanisms respectively (Peelle and Sommers, 2015).

The notion that an early integration mechanism increases auditory cortical sensitivity seems highly relevant in the context of speech-in-noise. Here, we demonstrated that neural tracking of audio-only speech in noise was significantly diminished at time lags between 0–95 ms, suggesting that auditory cortical sensitivity was reduced at an early stage of speech processing. Although the current data indicate that envelope tracking was restored by the addition of visual speech input at this early processing stage, because we include the entire head during the reconstruction analysis, it is difficult to say whether this is the result of increased auditory cortical sensitivity or rather contributions from multisensory areas such as STS or visual cortical areas. However, a theory that supports this notion of an early increase in auditory cortical sensitivity is that of cross-sensory phase-resetting of auditory cortex (Kayser et al., 2008, Schroeder et al., 2008, Mercier et al., 2015, Arnal et al., 2009, Lakatos et al., 2007). While such a mechanism can be difficult to reconcile in the context of extended vocalizations given that the time lag between visual and auditory speech is so variable (Schwartz and Savariaux, 2014), this can somewhat be explained by the temporal correspondence between the hierarchical organization of speech and that of the rhythmic oscillations in primary auditory cortex (Schroeder et al., 2008, Giraud and

Poeppel, 2012). While intuitively, it may seem more likely that auditory cortex would be primed by continuous visual input in a tonic manner, the idea of phasic crossmodal priming is supported by the fact that the temporal coherence between the A and V streams is critical for enhanced neural tracking during AV speech (Crosse et al., 2015a). This is also supported by accounts of enhanced phasic coordination across auditory and visual cortices for matched versus mismatched AV stimuli (Luo et al., 2010).

Evidence of a later integration stage that constrains lexical selection can also be found in numerous electrophysiological studies. Both TRF and ERP measures have revealed emergent multisensory interaction effects in the form of reduced component amplitude (Crosse et al., 2015a, Besle et al., 2004a, Bernstein et al., 2008, van Wassenhove et al., 2005). This reduction in cortical activation may well reflect a mechanism that constrains lexical computations based on the content of preceding visual information. Furthermore, the emergent interaction effect reported in Bernstein et al. (2008) was observed in left supramarginal and angular gyrus, in line with our left-dominant MSI effect. This left-bias could be explained by the fact that our data suggest that we integrate AV speech at a higher linguistic level in noise, and it has been suggested that speech processing becomes more left-lateralized further up the auditory hierarchy (Peelle, 2012). Both our single-lag analysis and temporal window analysis further suggest that integrating later temporal information contributes to AV speech processing. However, the most compelling evidence that is provided in favour of a late integration stage is the correspondence that was observed between the behavioural and neural measures at 220–250 ms. Given the likelihood that both of these measures reflect integration at the lexical-semantic level fits well with current views on the timecourse of such linguistic processing (Salmelin, 2007, Picton, 2013).

In summary, our results support the theory that visual speech helps increase auditory cortical sensitivity early on and constrains lexical processing of an acoustic utterance at a late computational stage. We contend that inverse effectiveness, which likely occurs as a result of multiple integration mechanisms, relies heavily on our ability to integrate crossmodal information over longer temporal windows during AV speech-in-noise.

7.5 Contributions and Future Directions

The present thesis has established an SI framework for investigating multisensory integration in the context of natural, continuous AV speech and, in doing so, yielded several significant insights into the neural basis of AV speech processing. Aside from furthering our understanding of how the human brain integrates AV speech, this naturalistic approach may yet prove useful in research with clinical populations in which altered multisensory processing has been reported, e.g., dyslexia (Hairston et al., 2005), ASD (Brandwein et al., 2013, Stevenson et al., 2014b), and schizophrenia (Ross et al., 2007b, Stekelenburg et al., 2013). In particular, neurodevelopmental disorders such as ASD, which need to be studied in young, pre-adolescent cohorts, would certainly benefit from a framework that facilitates the use of natural, engaging stimuli in an experimental context. Furthermore, impaired multisensory processing has been shown to be most pronounced in children with ASD for more complex stimuli such as AV speech (see section 1.1; Bebko et al., 2006). Future work could apply the SI framework developed in this thesis to study natural AV speech processing in children with ASD using AV audiobooks of children's fiction. From a basic scientific perspective, such work could enable researchers to investigate whether multisensory integration is impaired in ASD populations at an early or late stage of speech processing. This could not only further our understanding of the disorder, but inform better diagnosis and treatment of ASD.

Relating neural measures of multisensory integration to behavioural measures is an area that has gained attention due to its potential clinical utility. Previous work has tried to relate behavioural measures of multisensory integration, such as RT, to neural amplitude measures of integration (Stevenson et al., 2012a). A more recent clinical study reported that EEG indices of atypical multisensory processing were associated with ASD symptom severity (Brandwein et al., 2015). In Chapter 5, it was demonstrated that the neural index of multisensory integration, developed as part of this research work, was predictive of the multisensory gain in behaviour. This result, if replicable in children, would certainly have important implications in ASD research. Future work could look at refining this neural index of MSI and examine its utility as a biomarker for determining ASD symptom severity and/or monitoring developmental outcome in response to different therapeutic intervention strategies.

In Chapters 4 and 6, it was demonstrated that an estimate of the auditory speech envelope could be reconstructed from EEG recorded during silent lipreading with accuracy above chance level. These findings could have implications for the design of future BCI technologies that aim to decode internal speech from the user's neural recordings. As discussed in the previous chapter, decoding extended passages of covert speech has successfully been demonstrated using intracranial recordings such as ECoG (Martin et al., 2014). Other intracranial BCI approaches have sought to decode imagined speech at the level of phonemes, vowels and words (Guenther et al., 2009, Leuthardt et al., 2011, Kellis et al., 2010, Martin et al., 2016). The findings presented here indicate that EEG could provide a non-invasive, cost-effective solution to decoding imagined thoughts and could be further optimised by utilizing the natural statistics of visual speech input. Indeed, such technology would also have major implications for clinical research in populations that are unable to effectively communicate due to suffering from what's known as a 'locked-in syndrome', e.g., amyotrophic lateral sclerosis (Lou Gehrig's disease), traumatic brain injuries and spinal cord injuries. Moreover, the approach of speech decoding would provide a more naturalistic and user-friendly way for patients to communicate their thoughts compared to traditional EEG-based BCI methods that have primarily relied on discrete brain components that can only be elicited to specific target stimuli (Oken et al., 2014, Lesenfants et al., 2014, Combaz et al., 2013) – such methods can be tedious, time-consuming and ineffective (for a review, see Machado et al., 2010).

The research presented in this thesis highlights the utility of the mTRF Toolbox for research on questions relevant to AV speech processing and suggests potential applications to research on several clinical questions, including ASD and BCI technology. This thesis also reflects how the field of multisensory integration is expanding and demonstrates that MSI can be studied in a more naturalistic and ecologically valid manner. This is of great importance if we are to further our understanding of how the brain integrates multisensory information and, more generally, how it functions (or dysfunctions) in everyday situations.

References

ABRAMS, D. A., NICOL, T., ZECKER, S. & KRAUS, N. 2008. Right-hemisphere auditory cortex is dominant for coding syllable patterns in speech. *Journal of Neuroscience,* 28**,** 3958-3965.

ADANK, P. 2012. Design choices in imaging speech comprehension: an activation likelihood estimation (ALE) meta-analysis. *Neuroimage,* 63**,** 1601-1613.

AHISSAR, E., NAGARAJAN, S., AHISSAR, M., PROTOPAPAS, A., MAHNCKE, H. & MERZENICH, M. M. 2001. Speech comprehension is correlated with temporal response patterns recorded from auditory cortex. *Proceedings of the National Academy of Sciences,* 98**,** 13367-13372.

AIKEN, S. J. & PICTON, T. W. 2008. Human cortical responses to the speech envelope. *Ear and hearing,* 29**,** 139-157.

ALAIS, D., NEWELL, F. N. & MAMASSIAN, P. 2010. Multisensory processing in review: from physiology to behaviour. *Seeing and perceiving,* 23**,** 3-38.

APA, D.-A. P. A. 2013. Diagnostic and statistical manual of mental disorders. *Arlington: American Psychiatric Publishing*.

ARNAL, L. H., MORILLON, B., KELL, C. A. & GIRAUD, A. L. 2009. Dual Neural Routing of Visual Facilitation in Speech Processing. *Journal of Neuroscience,* 29**,** 13445-13453.

ARNAL, L. H., WYART, V. & GIRAUD, A. L. 2011. Transitions in neural oscillations reflect prediction errors generated in audiovisual speech. *Nature Neuroscience,* 14**,** 797-U164.

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

AVILLAC, M., DENEVE, S., OLIVIER, E., POUGET, A. & DUHAMEL, J.-R. 2005. Reference frames for representing visual and tactile locations in parietal cortex. *Nature neuroscience,* 8**,** 941-949.

BAART, M., STEKELENBURG, J. J. & VROOMEN, J. 2014. Electrophysiological evidence for speech-specific audiovisual integration. *Neuropsychologia,* 53**,** 115-121.

BARLOW, H. B. 1961. Possible principles underlying the transformation of sensory messages. *Sensory Communication***,** 217-234.

BARLOW, H. B. 1972. Single units and sensation: a neuron doctrine for perceptual psychology? *Perception***,** 795-8.

BARRACLOUGH, N. E., XIAO, D., BAKER, C., ORAM, M. W. & PERRETT, D. 2005. Integration of visual and auditory information by superior temporal sulcus neurons responsive to the sight of actions. *Cognitive Neuroscience, Journal of,* 17**,** 377-391.

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.

BEAR, M. F., CONNORS, B. W. & PARADISO, M. A. 2007. *Neuroscience,* Lippincott Williams & Wilkins.

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**,** 809-824.

BEBKO, J. M., WEISS, J. A., DEMARK, J. L. & GOMEZ, P. 2006. Discrimination of temporal synchrony in intermodal events by children with autism and children with developmental disabilities without autism. *Journal of Child Psychology and Psychiatry,* 47**,** 88-98.

BERMAN, A. L. 1961. Interaction of cortical responses to somatic and auditory stimuli in anterior ectosylvian gyrus of cat. *Journal of neurophysiology,* 24**,** 608-620.

BERNSTEIN, L. E., AUER, E. T., WAGNER, M. & PONTON, C. W. 2008. Spatiotemporal dynamics of audiovisual speech processing. *Neuroimage,* 39**,** 423-435.

BERNSTEIN, L. E., AUER JR, E. T. & MOORE, J. K. 2004a. Convergence or Association? *The handbook of multisensory processes***,** 203.

BERNSTEIN, L. E., AUER JR, E. T. & TAKAYANAGI, S. 2004b. Auditory speech detection in noise enhanced by lipreading. *Speech Communication,* 44**,** 5-18.

BERNSTEIN, L. E. & LIEBENTHAL, E. 2014. Neural pathways for visual speech perception. *Frontiers in neuroscience,* 8.

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,* 28**,** 14301-14310.

BESLE, J., FORT, A., DELPUECH, C. & GIARD, M. H. 2004a. Bimodal speech: early suppressive visual effects in human auditory cortex. *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.

BIALEK, W., RIEKE, F., VAN STEVENINCK, R. D. R. & WARLAND, D. 1991. Reading a neural code. *Science,* 252**,** 1854-1857.

BINNIE, C. A., MONTGOMERY, A. A. & JACKSON, P. L. 1974. Auditory and visual contributions to the perception of consonants. *Journal of speech, language, and hearing research,* 17**,** 619-630.

BLAMEY, P. J., COWAN, R. S., ALCANTARA, J. I., WHITFORD, L. A. & CLARK, G. M. 1989. Speech perception using combinations of auditory, visual, and tactile information. *Scientific publications, vol. 5, 1989-1990, no. 268*.

BOK, S. T. 1959. *Histonomy of the cerebral cortex*, Elsevier Pub. Co.

BRAIDA, L. D. 1991. Crossmodal integration in the identification of consonant segments. *The Quarterly Journal of Experimental Psychology,* 43**,** 647-677.

BRANDWEIN, A. B., FOXE, J. J., BUTLER, J. S., FREY, H.-P., BATES, J. C., SHULMAN, L. H. & MOLHOLM, S. 2015. Neurophysiological indices of atypical auditory processing and multisensory integration are associated with symptom severity in autism. *Journal of autism and developmental disorders,* 45**,** 230-244.

BRANDWEIN, A. B., FOXE, J. J., BUTLER, J. S., RUSSO, N. N., ALTSCHULER, T. S., GOMES, H. & MOLHOLM, S. 2013. The development of multisensory integration in high-functioning autism: high-density electrical mapping and psychophysical measures reveal impairments in the processing of audiovisual inputs. *Cerebral Cortex,* 23.

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. & LEWIS, J. 2004. Hemodynamic studies of audiovisual interaction. *The handbook of multisensory perception (eds GA Calvert, C. Spence & BE Stein)***,** 483-502.

CALVERT, G. A., BULLMORE, E. T., BRAMMER, M. J., CAMPBELL, R., WILLIAMS, S. C. R., MCGUIRE, P. K., WOODRUFF, P. W. R., IVERSON, S. D. & DAVID, A. S. 1997. Activation of auditory cortex during silent lipreading. *Science,* 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.

CAMPBELL, R., MACSWEENEY, M., SURGULADZE, S., CALVERT, G., MCGUIRE, P., SUCKLING, J., BRAMMER, M. J. & DAVID, A. S. 2001. Cortical substrates for the perception of face actions: an fMRI study of the specificity of activation for seen speech and for meaningless lower-face acts (gurning). *Cognitive Brain Research,* 12**,** 233-243.

CAPEK, C. M., BAVELIER, D., CORINA, D., NEWMAN, A. J., JEZZARD, P. & NEVILLE, H. J. 2004. The cortical organization of audio-visual sentence comprehension: an fMRI study at 4 Tesla. *Cognitive brain research,* 20**,** 111-119.

CHANDRASEKARAN, C., LEMUS, L. & GHAZANFAR, A. A. 2013. Dynamic faces speed up the onset of auditory cortical spiking responses during vocal detection. *Proceedings of the National Academy of Sciences,* 110**,** 4668-4677.

CHANDRASEKARAN, C., TRUBANOVA, A., STILLITTANO, S., CAPLIER, A. & GHAZANFAR, A. A. 2009. The natural statistics of audiovisual speech. *PLoS computational biology,* 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**,** 1428-1432.

CHOMSKY, N. & HALLE, M. 1968. The sound pattern of English.

COMBAZ, A., CHATELLE, C., ROBBEN, A., VANHOOF, G., GOELEVEN, A., THIJS, V., VAN HULLE, M. M. & LAUREYS, S. 2013. A comparison of two spelling brain-computer interfaces based on visual p3 and ssvep in locked-in syndrome. *PloS one,* 8**,** e73691.

COPPOLA, R. 1979. A system transfer function for visual evoked potentials. *Human Evoked Potentials.* Springer.

CORBY, J. C. & KOPELL, B. S. 1972. Differential contributions of blinks and vertical eye movements as artifacts in EEG recording. *Psychophysiology,* 9**,** 640-644.

CROSSE, M. J. 2011. *Nonlinear Regression Analysis for Assessing Human Auditory and Visual System Function.* Master of Science in Bioengineering, University College Dublin.

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*, 35**,** 14195-14204.

CROSSE, M. J., ELSHAFEI, H. A., FOXE, J. J. & LALOR, E. C. Investigating the temporal dynamics of auditory cortical activation to silent lipreading. Neural Engineering (NER), 2015 7th International IEEE/EMBS Conference on, 22–24 April 2015 2015b Montpellier. IEEE, 308-311.

CROSSE, M. J., O'SULLIVAN, J. A., POWER, A. J. & LALOR, E. C. The effects of attention and visual input on the representation of natural speech in EEG. Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE, 3-7 July 2013 2013. IEEE, 2800-2803.

DAVID, S. V. & GALLANT, J. L. 2005. Predicting neuronal responses during natural vision. *Network: Computation in Neural Systems,* 16**,** 239-260.

DAVID, S. V., MESGARANI, N. & SHAMMA, S. A. 2007. Estimating sparse spectro-temporal receptive fields with natural stimuli. *Network: Computation in Neural Systems,* 18**,** 191-212.

DAVIS, M. H. & GASKELL, M. G. 2009. A complementary systems account of word learning: neural and behavioural evidence. *Philosophical Transactions of the Royal Society B: Biological Sciences,* 364**,** 3773-3800.

DAVIS, P. A. 1939. Effects of acoustic stimuli on the waking human brain. *Journal of Neurophysiology,* 2**,** 494-499.

DAWSON, G., ROGERS, S., MUNSON, J., SMITH, M., WINTER, J., GREENSON, J., DONALDSON, A. & VARLEY, J. 2010. Randomized, controlled trial of an intervention for toddlers with autism: the Early Start Denver Model. *Pediatrics,* 125**,** e17-e23.

DE BOER, E. & KUYPER, P. 1968. Triggered correlation. *Biomedical Engineering, IEEE Transactions on***,** 169-179.

DE CHEVEIGNÉ, A. & PARRA, L. C. 2014. Joint decorrelation, a versatile tool for multichannel data analysis. *Neuroimage,* 98**,** 487-505.

DEPIREUX, D. A., SIMON, J. Z., KLEIN, D. J. & SHAMMA, S. A. 2001. Spectro-temporal response field characterization with dynamic ripples in ferret primary auditory cortex. *Journal of neurophysiology,* 85**,** 1220-1234.

DEWITT, I. & RAUSCHECKER, J. P. 2012. Phoneme and word recognition in the auditory ventral stream. *Proceedings of the National Academy of Sciences of the United States of America*, 109**,** 505-514.

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.

DI RUSSO, F., PITZALIS, S., SPITONI, G., APRILE, T., PATRIA, F., SPINELLI, D. & HILLYARD, S. A. 2005. Identification of the neural sources of the pattern-reversal VEP. *Neuroimage,* 24**,** 874-886.

DING, N., CHATTERJEE, M. & SIMON, J. Z. 2014. Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure. *Neuroimage,* 88**,** 41-46.

DING, N., MELLONI, L., ZHANG, H., TIAN, X. & POEPPEL, D. 2016. Cortical tracking of hierarchical linguistic structures in connected speech. *Nature Neuroscience,* 19**,** 158-164.

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. *Journal of Neuroscience,* 33**,** 5728-5735.

DING, N. & SIMON, J. Z. 2014. Cortical entrainment to continuous speech: functional roles and interpretations. *Frontiers in human neuroscience,* 8.

DONOHUE, S. E., DARLING, E. F. & MITROFF, S. R. 2012. Links between multisensory processing and autism. *Experimental brain research,* 222**,** 377-387.

DRIVER, J. & NOESSELT, T. 2008. Multisensory interplay reveals crossmodal influences on 'sensory-specific'brain regions, neural responses, and judgments. *Neuron,* 57**,** 11-23.

DRIVER, J. & SPENCE, C. 2000. Multisensory perception: beyond modularity and convergence. *Current Biology,* 10**,** R731-R735.

DRULLMAN, R., FESTEN, J. M. & PLOMP, R. 1994. Effect of temporal envelope smearing on speech reception. *The Journal of the Acoustical Society of America,* 95**,** 1053-1064.

DUNN, W., MYLES, B. S. & ORR, S. 2002. Sensory processing issues associated with Asperger syndrome: A preliminary investigation. *American Journal of Occupational Therapy,* 56**,** 97-102.

EGGERMONT, J., AERTSEN, A. & JOHANNESMA, P. 1983. Quantitative characterisation procedure for auditory neurons based on the spectro-temporal receptive field. *Hearing research,* 10**,** 167-190.

EHRET, G. & ROMAND, R. 1997. *The central auditory system*, Oxford University Press.

ELHILALI, M., MA, L., MICHEYL, C., OXENHAM, A. J. & SHAMMA, S. A. 2009. Temporal coherence in the perceptual organization and cortical representation of auditory scenes. *Neuron,* 61**,** 317-329.

ENGEL, A. K., FRIES, P. & SINGER, W. 2001. Dynamic predictions: oscillations and synchrony in top–down processing. *Nature Reviews Neuroscience,* 2**,** 704-716.

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

ERBER, N. P. 1971. Auditory and audiovisual reception of words in low-frequency noise by children with normal hearing and by children with impaired hearing. *Journal of Speech, Language, and Hearing Research,* 14**,** 496-512.

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

ESKELUND, K., TUOMAINEN, J. & ANDERSEN, T. S. 2011. Multistage audiovisual integration of speech: Dissociating identification and detection. *Experimental Brain Research,* 208**,** 447-457.

FALCHIER, A., CLAVAGNIER, S., BARONE, P. & KENNEDY, H. 2002. Anatomical evidence of Multimodal integration in primate striate cortex. *Journal of Neuroscience,* 22**,** 5749-5759.

FISHER, C. G. 1968. Confusions among visually perceived consonants. *Journal of Speech, Language, and Hearing Research,* 11**,** 796-804.

FOWLER, C. A. & DEKLE, D. J. 1991. Listening with eye and hand: cross-modal contributions to speech perception. *Journal of Experimental Psychology: Human Perception and Performance,* 17**,** 816.

FOXE, J. J. & MOLHOLM, S. 2009. Ten years at the multisensory forum: musings on the evolution of a field. *Brain topography,* 21**,** 149-154.

FOXE, J. J., MOLHOLM, S., DEL BENE, V. A., FREY, H.-P., RUSSO, N. N., BLANCO, D., SAINT-AMOUR, D. & ROSS, L. A. 2015. Severe multisensory speech integration deficits in high-functioning school-aged children with autism spectrum disorder (ASD) and their resolution during early adolescence. *Cerebral Cortex,* 25**,** 298-312.

FOXE, J. J. & SCHROEDER, C. E. 2005. The case for feedforward multisensory convergence during early cortical processing. *Neuroreport,* 16**,** 419-423.

FREEMAN, W. J., HOLMES, M. D., BURKE, B. C. & VANHATALO, S. 2003. Spatial spectra of scalp EEG and EMG from awake humans. *Clinical Neurophysiology,* 114**,** 1053-1068.

FREY, H.-P., KELLY, S. P., LALOR, E. C. & FOXE, J. J. 2010. Early spatial attentional modulation of inputs to the fovea. *The Journal of Neuroscience,* 30**,** 4547-4551.

FREY, H. P., MOLHOLM, S., LALOR, E. C., RUSSO, N. N. & FOXE, J. J. 2013. Atypical cortical representation of peripheral visual space in children with an autism spectrum disorder. *European Journal of Neuroscience,* 38**,** 2125-2138.

FRIDRIKSSON, J., MOSS, J., DAVIS, B., BAYLIS, G. C., BONILHA, L. & RORDEN, C. 2008. Motor speech perception modulates the cortical language areas. *Neuroimage,* 41**,** 605-613.

FRISTON, K. 2005. A theory of cortical responses. *Philosophical Transactions of the Royal Society B-Biological Sciences,* 360**,** 815-836.

GEVINS, A., LE, J., BRICKETT, P., REUTTER, B. & DESMOND, J. 1991. Seeing through the skull: advanced EEGs use MRIs to accurately measure cortical activity from the scalp. *Brain Topography,* 4**,** 125–131.

GHAZANFAR, A. A., MAIER, J. X., HOFFMAN, K. L. & LOGOTHETIS, N. K. 2005. Multisensory integration of dynamic faces and voices in rhesus monkey auditory cortex. *The Journal of Neuroscience,* 25**,** 5004-5012.

GHAZANFAR, A. A. & SCHROEDER, C. E. 2006. Is neocortex essentially multisensory? *Trends in cognitive sciences,* 10**,** 278-285.

GIRAUD, A.-L. & POEPPEL, D. 2012. Cortical oscillations and speech processing: emerging computational principles and operations. *Nature neuroscience,* 15**,** 511-517.

GONÇALVES, N. R., WHELAN, R., FOXE, J. J. & LALOR, E. C. 2014. Towards obtaining spatiotemporally precise responses to continuous sensory stimuli in humans: A general linear modeling approach to EEG. *NeuroImage,* 97**,** 196-205.

GONCHAROVA, I., MCFARLAND, D. J., VAUGHAN, T. M. & WOLPAW, J. R. 2003. EMG contamination of EEG: spectral and topographical characteristics. *Clinical Neurophysiology,* 114**,** 1580-1593.

GRANT, K. W. 2001. The effect of speechreading on masked detection thresholds for filtered speech. *The Journal of the Acoustical Society of America,* 109**,** 2272-2275.

GRANT, K. W., GREENBERG, S., POEPPEL, D. & VAN WASSENHOVE, V. Effects of spectro-temporal asynchrony in auditory and auditory-visual speech processing. Seminars in Hearing, 2004. 241-255.

GRANT, K. W. & SEITZ, P. F. 1998. Measures of auditory–visual integration in nonsense syllables and sentences. *The Journal of the Acoustical Society of America,* 104**,** 2438-2450.

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

GRANT, K. W., TUFTS, J. B. & GREENBERG, S. 2007. Integration efficiency for speech perception within and across sensory modalities by normal-hearing and hearing-impaired individuals. *Journal of the Acoustical Society of America,* 121**,** 1164-1176.

GRANT, K. W. & WALDEN, B. E. 1996. Evaluating the articulation index for auditory–visual consonant recognition. *The Journal of the Acoustical Society of America,* 100**,** 2415-2424.

GRANT, K. W., WALDEN, B. E. & SEITZ, P. F. 1998. Auditory-visual speech recognition by hearing-impaired subjects: Consonant recognition, sentence recognition, and auditory-visual integration. *The Journal of the Acoustical Society of America,* 103**,** 2677-2690.

GREEN, K. P. & KUHL, P. K. 1989. The role of visual information in the processing of. *Perception & Psychophysics,* 45**,** 34-42.

GREENWOOD, D. D. 1990. A cochlear frequency-position function for several species—29 years later. *The Journal of the Acoustical Society of America,* 87**,** 2592-2605.

GUENTHER, F. H., BRUMBERG, J. S., WRIGHT, E. J., NIETO-CASTANON, A., TOURVILLE, J. A., PANKO, M., LAW, R., SIEBERT, S. A., BARTELS, J. L. & ANDREASEN, D. S. 2009. A wireless brain-machine interface for real-time speech synthesis. *PloS one,* 4**,** e8218.

HAIRSTON, W. D., BURDETTE, J. H., FLOWERS, D. L., WOOD, F. B. & WALLACE, M. T. 2005. Altered temporal profile of visual-auditory multisensory interactions in dyslexia. *Experimental Brain Research,* 166**,** 474-480.

HANDY, T. C. 2005. *Event-related potentials: A methods handbook*, The MIT Press.

HARI, R. & SALMELIN, R. 1997. Human cortical oscillations: a neuromagnetic view through the skull. *Trends in neurosciences,* 20**,** 44-49.

HAUFE, S., MEINECKE, F., GÖRGEN, K., DÄHNE, 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.

HICKOK, G. & POEPPEL, D. 2007. The cortical organization of speech processing. *Nature Reviews Neuroscience,* 8**,** 393-402.

HOLMES, N. P. 2009. The principle of inverse effectiveness in multisensory integration: some statistical considerations. *Brain topography,* 21**,** 168-176.

HOUTGAST, T. & STEENEKEN, H. J. 1985. A review of the MTF concept in room acoustics and its use for estimating speech intelligibility in auditoria. *The Journal of the Acoustical Society of America,* 77**,** 1069-1077.

HOWARD, M. F. & POEPPEL, D. 2010. Discrimination of speech stimuli based on neuronal response phase patterns depends on acoustics but not comprehension. *Journal of neurophysiology,* 104**,** 2500-2511.

IRINO, T. & PATTERSON, R. D. 2006. A dynamic compressive gammachirp auditory filterbank. *Audio, Speech, and Language Processing, IEEE Transactions on,* 14**,** 2222-2232.

JASPER, H. H. 1958. The ten twenty electrode system of the international federation. *Electroencephalography and clinical neurophysiology,* 10**,** 371-375.

JIANG, J., ALWAN, A., KEATING, P. A., AUER, E. T. & BERNSTEIN, L. E. 2002. On the relationship between face movements, tongue movements, and speech acoustics. *EURASIP Journal on Applied Signal Processing,* 11**,** 1174-1188.

JIANG, J. & BERNSTEIN, L. E. 2011. Psychophysics of the McGurk and other audiovisual speech integration effects. *Journal of Experimental Psychology: Human Perception and Performance,* 37**,** 1193-1209.

JONES, J. P. & PALMER, L. A. 1987. The two-dimensional spatial structure of simple receptive fields in cat striate cortex. *Journal of neurophysiology,* 58**,** 1187-1211.

JONES, M. R., JOHNSTON, H. M. & PUENTE, J. 2006. Effects of auditory pattern structure on anticipatory and reactive attending. *Cognitive psychology,* 53**,** 59-96.

KAYSER, C. & LOGOTHETIS, N. K. 2007. Do early sensory cortices integrate cross-modal information? *Brain structure and function,* 212**,** 121-132.

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.

KAYSER, C., LOGOTHETIS, N. K. & PANZERI, S. 2010. Visual enhancement of the information representation in auditory cortex. *Current Biology,* 20**,** 19-24.

KAYSER, C., PETKOV, C. I., AUGATH, M. & LOGOTHETIS, N. K. 2007. Functional imaging reveals visual modulation of specific fields in auditory cortex. *The Journal of neuroscience,* 27**,** 1824-1835.

KAYSER, C., PETKOV, C. I. & LOGOTHETIS, N. K. 2008. Visual modulation of neurons in auditory cortex. *Cerebral Cortex,* 18**,** 1560-1574.

KAYSER, C., PETKOV, C. I. & LOGOTHETIS, N. K. 2009. Multisensory interactions in primate auditory cortex: fMRI and electrophysiology. *Hearing research,* 258**,** 80-88.

KELLIS, S., MILLER, K., THOMSON, K., BROWN, R., HOUSE, P. & GREGER, B. 2010. Decoding spoken words using local field potentials recorded from the cortical surface. *Journal of neural engineering,* 7**,** 056007.

KENT, J. L. 2010. *Psychedelic information theory: Shamanism in the age of reason*, PIT Press/Supermassive.

KINCHLA, R. 1974. Detecting target elements in multielement arrays: A confusability model. *Perception & Psychophysics,* 15**,** 149-158.

KLUCHAREV, V., MOTTONEN, R. & SAMS, M. 2003. Electrophysiological indicators of phonetic and non-phonetic multisensory interactions during audiovisual speech perception. *Cognitive Brain Research,* 18**,** 65-75.

KRASOULIS, A., VIJAYAKUMAR, S. & NAZARPOUR, K. Evaluation of regression methods for the continuous decoding of finger movement from surface EMG and accelerometry. Neural Engineering (NER), 2015 7th International IEEE/EMBS Conference on, 2015. IEEE, 631-634.

KRUSKAL, J. B. & WISH, M. 1978. *Multidimensional scaling*, Sage.

KUWABARA, H. Acoustic properties of phonemes in continuous speech for different speaking rate. Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on, 1996. IEEE, 2435-2438.

LAKATOS, P., CHEN, C. M., O'CONNELL, M. N., MILLS, A. & SCHROEDER, C. E. 2007. Neuronal oscillations and multisensory interaction in primary auditory cortex. *Neuron,* 53**,** 279-292.

LAKATOS, P., SHAH, A. S., KNUTH, K. H., ULBERT, I., KARMOS, G. & SCHROEDER, C. E. 2005. An oscillatory hierarchy controlling neuronal excitability and stimulus processing in the auditory cortex. *Journal of neurophysiology,* 94**,** 1904-1911.

LALOR, E. C., DE SANCTIS, P., KRAKOWSKI, M. I. & FOXE, J. J. 2012. Visual sensory processing deficits in schizophrenia: is there anything to the magnocellular account? *Schizophrenia research,* 139**,** 246-252.

LALOR, E. C. & FOXE, J. J. 2010. Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. *European Journal of Neuroscience,* 31**,** 189-193.

LALOR, E. C., KELLY, S. P., PEARLMUTTER, B. A., REILLY, R. B. & FOXE, J. J. 2007. Isolating endogenous visuo-spatial attentional effects using the novel visual-evoked spread spectrum analysis (VESPA) technique. *European Journal of Neuroscience,* 26**,** 3536-3542.

LALOR, E. C., PEARLMUTTER, B. A., REILLY, R. B., MCDARBY, G. & FOXE, J. J. 2006. The VESPA: a method for the rapid estimation of a visual evoked potential. *Neuroimage,* 32**,** 1549-1561.

LALOR, E. C., POWER, A. J., REILLY, R. B. & FOXE, J. J. 2009. Resolving precise temporal processing properties of the auditory system using continuous stimuli. *Journal of Neurophysiology,* 102**,** 349-359.

LALOR, E. C., YEAP, S., REILLY, R. B., PEARLMUTTER, B. A. & FOXE, J. J. 2008. Dissecting the cellular contributions to early visual sensory processing deficits in schizophrenia using the VESPA evoked response. *Schizophrenia Research,* 98**,** 256-264.

LARGE, E. W. & JONES, M. R. 1999. The dynamics of attending: How people track time-varying events. *Psychological Review,* 106**,** 119-159.

LEEKAM, S. R., NIETO, C., LIBBY, S. J., WING, L. & GOULD, J. 2007. Describing the sensory abnormalities of children and adults with autism. *Journal of autism and developmental disorders,* 37**,** 894-910.

LEHMANN, D. & SKRANDIES, W. 1980. Reference-free identification of components of checkerboard-evoked multichannel potential fields. *Electroencephalography and Clinical Neurophysiology,* 48**,** 609-621.

LESENFANTS, D., HABBAL, D., LUGO, Z., LEBEAU, M., HORKI, P., AMICO, E., POKORNY, C., GOMEZ, F., SODDU, A. & MÜLLER-PUTZ, G. 2014. An independent SSVEP-based brain–computer interface in locked-in syndrome. *Journal of neural engineering,* 11**,** 035002.

LEUTHARDT, E. C., GAONA, C., SHARMA, M., SZRAMA, N., ROLAND, J., FREUDENBERG, Z., SOLIS, J., BRESHEARS, J. & SCHALK, G. 2011. Using the electrocorticographic speech network to control a brain–computer interface in humans. *Journal of neural engineering,* 8**,** 036004.

LUO, H., LIU, Z. X. & POEPPEL, D. 2010. Auditory Cortex Tracks Both Auditory and Visual Stimulus Dynamics Using Low-Frequency Neuronal Phase Modulation. *Plos Biology,* 8.

LUO, H. & POEPPEL, D. 2007. Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron,* 54**,** 1001-1010.

MACHADO, S., ARAÚJO, F., PAES, F., VELASQUES, B., CUNHA, M., BUDDE, H., BASILE, L. F., ANGHINAH, R., ARIAS-CARRIÓN, O. & CAGY, M. 2010. EEG-based brain-computer interfaces: an overview of basic concepts and clinical applications in neurorehabilitation. *Reviews in the neurosciences,* 21**,** 451-468.

MACHENS, C. K., WEHR, M. S. & ZADOR, A. M. 2004. Linearity of cortical receptive fields measured with natural sounds. *The Journal of neuroscience,* 24**,** 1089-1100.

MAIER, J. X., CHANDRASEKARAN, C. & GHAZANFAR, A. A. 2008. Integration of bimodal looming signals through neuronal coherence in the temporal lobe. *Current Biology,* 18**,** 963-968.

MAIER, J. X., DI LUCA, M. & NOPPENEY, U. 2011. Audiovisual asynchrony detection in human speech. *Journal of Experimental Psychology: Human Perception and Performance,* 37**,** 245.

MARCO, E. J., HINKLEY, L. B., HILL, S. S. & NAGARAJAN, S. S. 2011. Sensory processing in autism: a review of neurophysiologic findings. *Pediatric Research,* 69**,** 48R-54R.

MARMARELIS, P. & MARMARELIS, V. 1978. Analysis of physiological systems: the white-noise approach. Plenum Press, New York.

MARMARELIS, V. Z. 2004. *Nonlinear dynamic modeling of physiological systems*, John Wiley & Sons.

MARTIN, S., BRUNNER, P., HOLDGRAF, C., HEINZE, H.-J., CRONE, N. E., RIEGER, J., SCHALK, G., KNIGHT, R. T. & PASLEY, B. N. 2014. Decoding spectrotemporal features of overt and covert speech from the human cortex. *Frontiers in Neuroengineering,* 7.

MARTIN, S., BRUNNER, P., ITURRATE, I., MILLÁN, J. D. R., SCHALK, G., KNIGHT, R. T. & PASLEY, B. N. 2016. Word pair classification during imagined speech using direct brain recordings. *Scientific reports,* 6.

MARTINI, F. H. & NATH, J. L. 2009. *Fundamentals of Anatomy and Physiology,* San Francisco, Pearson Education.

MASSARO, D. W. 1999. Speechreading: illusion or window into pattern recognition. *Trends in Cognitive Sciences,* 3**,** 310-317.

MASSARO, D. W., COHEN, M. M. & SMEELE, P. M. 1995. Cross-linguistic comparisons in the integration of visual and auditory speech. *Memory & Cognition,* 23**,** 113-131.

MCCORMICK, B. 1979. Audio-visual discrimination of speech*. *Clinical Otolaryngology & Allied Sciences,* 4**,** 355-361.

MCDEVITT, N., GALLAGHER, L. & REILLY, R. B. 2015. Autism Spectrum Disorder (ASD) and Fragile X Syndrome (FXS): Two Overlapping Disorders Reviewed through Electroencephalography—What Can be Interpreted from the Available Information? *Brain sciences,* 5**,** 92-117.

MCGURK, H. & MACDONALD, J. 1976. Hearing lips and seeing voices. *Nature,* 264**,** 746-748.

MEGNIN, O., FLITTON, A., RG JONES, C., DE HAAN, M., BALDEWEG, T. & CHARMAN, T. 2012. Audiovisual speech integration in autism spectrum disorders: ERP evidence for atypicalities in lexical-semantic processing. *Autism Research,* 5**,** 39-48.

MERCIER, M. R., MOLHOLM, S., FIEBELKORN, I. C., BUTLER, J. S., SCHWARTZ, T. H. & FOXE, J. J. 2015. Neuro-Oscillatory Phase Alignment Drives Speeded Multisensory Response Times: An Electro-Corticographic Investigation. *The Journal of Neuroscience,* 35**,** 8546-8557.

MEREDITH, M. A., NEMITZ, J. W. & STEIN, B. E. 1987. Determinants of multisensory integration in superior colliculus neurons. I. Temporal factors. *The Journal of neuroscience,* 7**,** 3215-3229.

MEREDITH, M. A. & STEIN, B. E. 1983. Interactions among converging sensory inputs in the superior colliculus. *Science,* 221**,** 389-391.

MEREDITH, M. A. & STEIN, B. E. 1985. Descending efferents from the superior colliculus relay integrated multisensory information. *Science,* 227**,** 657-659.

MEREDITH, M. A. & STEIN, B. E. 1986a. Spatial factors determine the activity of multisensory neurons in cat superior colliculus. *Brain Research,* 365**,** 350-354.

MEREDITH, M. A. & STEIN, B. E. 1986b. Visual, auditory, and somatosensory convergence on cells in superior colliculus results in multisensory integration. *Journal of neurophysiology,* 56**,** 640-662.

MESGARANI, N. & CHANG, E. F. 2012. Selective cortical representation of attended speaker in multi-talker speech perception. *Nature,* 485**,** 233-236.

MESGARANI, N., CHEUNG, C., JOHNSON, K. & CHANG, E. F. 2014. Phonetic feature encoding in human superior temporal gyrus. *Science,* 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.

MILLER, G. A., HEISE, G. A. & LICHTEN, W. 1951. The intelligibility of speech as a function of the context of the test materials. *Journal of experimental psychology,* 41**,** 329.

MILLER, G. A. & NICELY, P. E. 1955. An analysis of perceptual confusions among some English consonants. *The Journal of the Acoustical Society of America,* 27**,** 338-352.

MILLER, J. 1982. Divided attention: Evidence for coactivation with redundant signals. *Cognitive psychology,* 14**,** 247-279.

MILLER, L. M. & D'ESPOSITO, M. 2005. Perceptual fusion and stimulus coincidence in the cross-modal integration of speech. *Journal of Neuroscience,* 25**,** 5884-5893.

MIRKOVIC, B., DEBENER, S., JAEGER, M. & DE VOS, M. 2015. Decoding the attended speech stream with multi-channel EEG: implications for online, daily-life applications. *Journal of neural engineering,* 12**,** 046007.

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.

MÖTTÖNEN, R., KRAUSE, C. M., TIIPPANA, K. & SAMS, M. 2002. Processing of changes in visual speech in the human auditory cortex. *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.

MÖTTÖNEN, R. & WATKINS, K. E. 2009. Motor representations of articulators contribute to categorical perception of speech sounds. *The Journal of Neuroscience,* 29**,** 9819-9825.

MUNHALL, K. G., JONES, J. A., CALLAN, D. E., KURATATE, T. & VATIKIOTIS-BATESON, E. 2004a. Visual prosody and speech intelligibility - Head movement improves auditory speech perception. *Psychological Science,* 15**,** 133-137.

MUNHALL, K. G., JONES, J. A., CALLAN, D. E., KURATATE, T. & VATIKIOTIS-BATESON, E. 2004b. Visual prosody and speech intelligibility head movement improves auditory speech perception. *Psychological science,* 15**,** 133-137.

MURPHY, J. W., KELLY, S. P., FOXE, J. J. & LALOR, E. C. 2012. Isolating early cortical generators of visual-evoked activity: a systems identification approach. *Experimental brain research,* 220**,** 191-199.

MURRAY, M. M., BRUNET, D. & MICHEL, C. M. 2008. Topographic ERP analyses: a step-by-step tutorial review. *Brain topography,* 20**,** 249-264.

NATH, A. R. & BEAUCHAMP, M. S. 2011. Dynamic changes in superior temporal sulcus connectivity during perception of noisy audiovisual speech. *The Journal of Neuroscience,* 31**,** 1704-1714.

NEELY, K. K. 1956. Effect of visual factors on the intelligibility of speech. *The Journal of the Acoustical Society of America,* 28**,** 1275-1277.

NELKEN, I. 2008. Processing of complex sounds in the auditory system. *Current Opinion in Neurobiology,* 18**,** 413-417.

NOBRE, A. C., CORREA, A. & COULL, J. T. 2007. The hazards of time. *Current opinion in neurobiology,* 17**,** 465-470.

NOBRE, K. & COULL, J. T. 2010. *Attention and time,* Oxford University Press.

NOCK, H. J., IYENGAR, G. & NETI, C. Assessing face and speech consistency for monologue detection in video. Proceedings of the tenth ACM international conference on Multimedia, December 2002 Juan-les-Pins. ACM, 303-306.

NUNEZ, P. 1995. Quantitative states of neocortex. *Neocortical dynamics and human EEG rhythms***,** 3-67.

NUNEZ, P. L. & SRINIVASAN, R. 2006. *Electric fields of the brain: the neurophysics of EEG,* Oxford university press.

O'SULLIVAN, J. A., CROSSE, M. J., POWER, A. J. & LALOR, E. C. 2013. The effects of attention and visual input on the representation of natural speech in

EEG. Conference Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Jul 2013. IEEE, 2800–2803.

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. 2015. Attentional selection in a cocktail party environment can be decoded from single-trial EEG. *Cerebral Cortex,* 25**,** 1697-1706.

O’NEILL, J. J. 1954. Contributions of the visual components of oral symbols to speech comprehension. *Journal of Speech and Hearing Disorders,* 19**,** 429-439.

OJANEN, V., MÖTTÖNEN, R., PEKKOLA, J., JÄÄSKELÄINEN, I. P., JOENSUU, R., AUTTI, T. & SAMS, M. 2005. Processing of audiovisual speech in Broca's area. *Neuroimage,* 25**,** 333-338.

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**,** 68-959.

OKEN, B. S., ORHAN, U., ROARK, B., ERDOGMUS, D., FOWLER, A., MOONEY, A., PETERS, B., MILLER, M. & FRIED-OKEN, M. B. 2014. Brain–Computer Interface With Language Model–Electroencephalography Fusion for Locked-In Syndrome. *Neurorehabilitation and neural repair,* 28**,** 387-394.

PAPANICOLAOU, A. C. 1998. *Fundamentals of functional brain imaging: A guide to the methods and their applications to psychology and behavioral neuroscience,* CRC Press.

PARSONS, T. W. 1987. *Voice and speech processing,* McGraw-Hill College.

PASLEY, B. N., DAVID, S. V., MESGARANI, N., FLINKER, A., SHAMMA, S. A., CRONE, N. E., KNIGHT, R. T. & CHANG, E. F. 2012. Reconstructing Speech from Human Auditory Cortex. *Plos Biology,* 10.

PEELLE, J. E. 2012. The hemispheric lateralization of speech processing depends on what “speech” is: a hierarchical perspective. *Frontiers in human neuroscience,* 6.

PEELLE, J. E. & DAVIS, M. H. 2012. Neural oscillations carry speech rhythm through to comprehension. *Frontiers in psychology,* 3.

PEELLE, J. E., GROSS, J. & DAVIS, M. H. 2013. Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. *Cerebral cortex,* 23**,** 1378-1387.

PEELLE, J. E. & SOMMERS, M. S. 2015. Prediction and constraint in audiovisual speech perception. *Cortex,* 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.

PERRODIN, C., KAYSER, C., LOGOTHETIS, N. K. & PETKOV, C. I. 2015. Natural asynchronies in audiovisual communication signals regulate neuronal multisensory interactions in voice-sensitive cortex. *Proceedings of the National Academy of Sciences,* 112**,** 273-278.

PICTON, T. 2013. Hearing in time: evoked potential studies of temporal processing. *Ear and hearing,* 34**,** 385-401.

PICTON, T. W., HILLYARD, S. A., KRAUSZ, H. I. & GALAMBOS, R. 1974. Human auditory evoked potentials. I: Evaluation of components. *Electroencephalography and clinical neurophysiology,* 36**,** 179-190.

PILLING, M. 2009. Auditory Event-Related Potentials (ERPs) in Audiovisual Speech Perception. *Journal of Speech, Language, and Hearing Research,* 52**,** 1073-1081.

POEPPEL, D. 2003. The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric sampling in time'. *Speech Communication,* 41**,** 245-255.

POWER, A. J., FOXE, J. J., FORDE, E. J., REILLY, R. B. & LALOR, E. C. 2012. At what time is the cocktail party? A late locus of selective attention to natural speech. *European Journal of Neuroscience,* 35**,** 1497-1503.

POWER, A. J., REILLY, R. B. & LALOR, E. C. Comparing linear and quadratic models of the human auditory system using EEG. Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE, 2011a. IEEE, 4171-4174.

POWER, A. J., REILLY, R. B. & LALOR, E. C. 2011b. Comparison of Linear and Quadratic Modelling of the Human Auditory System Using a System Identification Approach.

PURVES, D., AUGUSTINE, G., FITZPATRICK, D., HALL, W., LAMANTIA, A., MCNAMARA, J. & WHITE, L. 2008. *Neuroscience,* Sunderland, MA, USA, Sinauer Associates, Inc.

RAAB, D. H. 1962. Statisical facilitation of simple reaction times. *Transactions of the New York Academy of Sciences,* 24**,** 574-590.

RABINER, L. R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE,* 77**,** 257-286.

RAHNE, T., DEIKE, S., SELEZNEVA, E., BROSCH, M., KÖNIG, R., SCHEICH, H., BÖCKMANN, M. & BRECHMANN, A. 2008. A multilevel and cross-modal approach towards neuronal mechanisms of auditory streaming. *Brain research,* 1220**,** 118-131.

REISBERG, D., MCLEAN, J. & GOLDFIELD, A. 1987. Easy to hear but hard to understand: A lip-reading advantage with intact auditory stimuli. *In:* DODD, B. & CAMPBELL, R. (eds.) *Hearing by eye: The psychology of lip-reading.* Lawrence Erlbaum Associates.

REMEZ, R. 2005. Three puzzles of multimodal speech perception. *Audiovisual speech***,** 12-19.

RICE, H. 2009. *General Principles of Acoustics,* Hertfordshire.

RIEKE, F., BODNAR, D. & BIALEK, W. 1995. Naturalistic stimuli increase the rate and efficiency of information transmission by primary auditory afferents. *Proceedings of the Royal Society of London B: Biological Sciences,* 262**,** 259-265.

RIJSBERGEN, C. & JOOST, C. K. 1979. Information Retrieval. Butterworth. Heinemann.

RINGACH, D. & SHAPLEY, R. 2004. Reverse correlation in neurophysiology. *Cognitive Science,* 28**,** 147-166.

RINGACH, D. L., SAPIRO, G. & SHAPLEY, R. 1997. A subspace reverse-correlation technique for the study of visual neurons. *Vision research,* 37**,** 2455-2464.

ROBINSON, D. J. & HAWKSFORD, M. J. Time-domain auditory model for the assessment of high-quality coded audio. Audio Engineering Society Convention 107, 1999. Audio Engineering Society.

ROSEN, S. 1992. Temporal information in speech: acoustic, auditory and linguistic aspects. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences,* 336**,** 367-373.

ROSENBLUM, L. D. 2005. Primacy of multimodal speech perception. *Handbook of speech perception***,** 51-78.

ROSENFELDER, I., FRUEHWALD, J., EVANINI, K. & YUAN, J. 2011. *FAVE (Forced Alignment and Vowel Extraction) Program Suite* [Online]. Available: http://fave.ling.upenn.edu [Accessed].

ROSS, L. A., SAINT-AMOUR, D., LEAVITT, V. M., JAVITT, D. C. & FOXE, J. J. 2007a. Do you see what I am saying? Exploring visual enhancement of speech comprehension in noisy environment. *Cerebral Cortex,* 17**,** 1147-1153.

ROSS, L. A., SAINT-AMOUR, D., LEAVITT, V. M., MOLHOLM, S., JAVITT, D. C. & FOXE, J. J. 2007b. Impaired multisensory processing in schizophrenia: Deficits in the visual enhancement of speech comprehension under noisy environmental conditions. *Schizophrenia Research,* 97**,** 173-183.

ROWLAND, B. A., QUESSY, S., STANFORD, T. R. & STEIN, B. E. 2007. Multisensory integration shortens physiological response latencies. *The Journal of neuroscience,* 27**,** 5879-5884.

RUSSO, N., FOXE, J. J., BRANDWEIN, A. B., ALTSCHULER, T., GOMES, H. & MOLHOLM, S. 2010. Multisensory processing in children with autism: high-density electrical mapping of auditory–somatosensory integration. *Autism Research,* 3**,** 253-267.

SAKKALIS, V., CASSAR, T., ZERVAKIS, M., GIURCANEANU, C. D., BIGAN, C., MICHELOYANNIS, S., CAMILLERI, K. P., FABRI, S. G., KARAKONSTANTAKI, E. & MICHALOPOULOS, K. 2010. A decision support framework for the discrimination of children with controlled epilepsy based on EEG analysis. *Journal of neuroengineering and rehabilitation,* 7**,** 24.

SALMELIN, R. 2007. Clinical neurophysiology of language: The MEG approach. *Clinical Neurophysiology,* 118**,** 237-254.

SAMS, M., AULANKO, R., HÄMÄLÄINEN, 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.

SAUSENG, P., KLIMESCH, W., GRUBER, W., HANSLMAYR, S., FREUNBERGER, R. & DOPPELMAYR, M. 2007. Are event-related potential components generated by phase resetting of brain oscillations? A critical discussion. *Neuroscience,* 146**,** 1435-1444.

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. 2010. A reanalysis of McGurk data suggests that audiovisual fusion in speech perception is subject-dependent. *The Journal of the Acoustical Society of America,* 127**,** 1584-1594.

SCHWARTZ, J.-L., BERTHOMMIER, F. & SAVARIAUX, C. 2004. Seeing to hear better: evidence for early audio-visual interactions in speech identification. *Cognition,* 93**,** B69-B78.

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.

SEKIYAMA, K. & TOHKURA, Y. I. 1991. McGurk effect in non-English listeners: Few visual effects for Japanese subjects hearing Japanese syllables of high auditory intelligibility. *The Journal of the Acoustical Society of America,* 90**,** 1797-1805.

SENKOWSKI, D., SCHNEIDER, T. R., FOXE, J. J. & ENGEL, A. K. 2008. Crossmodal binding through neural coherence: implications for multisensory processing. *Trends in neurosciences,* 31**,** 401-409.

SHAMMA, S. A., ELHILALI, M. & MICHEYL, C. 2011. Temporal coherence and attention in auditory scene analysis. *Trends in Neurosciences,* 34**,** 114-123.

SHANNON, R. V., ZENG, F.-G., KAMATH, V., WYGONSKI, J. & EKELID, M. 1995. Speech recognition with primarily temporal cues. *Science,* 270**,** 303-304.

SKIPPER, J. I., NUSBAUM, H. C. & SMALL, S. L. 2005. Listening to talking faces: motor cortical activation during speech perception. *Neuroimage,* 25**,** 76-89.

SKOKAUSKAS, N. & GALLAGHER, L. 2012. Mental health aspects of autistic spectrum disorders in children. *Journal of Intellectual Disability Research,* 56**,** 248-257.

SMITH, L. & GASSER, M. 2005. The development of embodied cognition: Six lessons from babies. *Artificial Life,* 11**,** 13-29.

STANLEY, G. B., LI, F. F. & DAN, Y. 1999. Reconstruction of natural scenes from ensemble responses in the lateral geniculate nucleus. *The Journal of Neuroscience,* 19**,** 8036-8042.

STEIN, B. E. & MEREDITH, M. A. 1993. *The merging of the senses*, The MIT Press.

STEIN, B. E. & STANFORD, T. R. 2008. Multisensory integration: current issues from the perspective of the single neuron. *Nature Reviews Neuroscience,* 9**,** 255-266.

STEIN, B. E., STANFORD, T. R. & ROWLAND, B. A. 2014. Development of multisensory integration from the perspective of the individual neuron. *Nature Reviews Neuroscience,* 15**,** 520-535.

STEIN, B. E. & WALLACE, M. T. 1996. Comparisons of cross-modality integration in midbrain and cortex. *Progress in brain research,* 112**,** 289-299.

STEKELENBURG, J. J., MAES, J. P., VAN GOOL, A. R., SITSKOORN, M. & VROOMEN, J. 2013. Deficient multisensory integration in schizophrenia: An event-related potential study. *Schizophrenia research,* 147**,** 253-261.

STEKELENBURG, J. J. & VROOMEN, J. 2007. Neural correlates of multisensory integration of ecologically valid audiovisual events. *Journal of Cognitive Neuroscience,* 19**,** 1964-1973.

STERIADE, M., GLOOR, P., LLINAS, R., DA SILVA, F. L. & MESULAM, M.-M. 1990. Basic mechanisms of cerebral rhythmic activities. *Electroencephalography and clinical neurophysiology,* 76**,** 481-508.

STEVENSON, R. A., BUSHMAKIN, M., KIM, S., WALLACE, M. T., PUCE, A. & JAMES, T. W. 2012a. Inverse effectiveness and multisensory interactions in visual event-related potentials with audiovisual speech. *Brain topography,* 25**,** 308-326.

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. 2014a. 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.

STEVENSON, R. A., SIEMANN, J. K., WOYNAROSKI, T. G., SCHNEIDER, B. C., EBERLY, H. E., CAMARATA, S. M. & WALLACE, M. T. 2014b. Evidence for diminished multisensory integration in autism spectrum disorders. *Journal of autism and developmental disorders,* 44**,** 3161-3167.

STEVENSON, R. A., ZEMTSOV, R. K. & WALLACE, M. T. 2012b. Individual differences in the multisensory temporal binding window predict susceptibility to audiovisual illusions. *Journal of Experimental Psychology: Human Perception and Performance,* 38**,** 1517.

SUMBY, W. H. & POLLACK, I. 1954. Visual contribution to speech intelligibility in noise. *Journal of the Acoustical Society of America,* 26**,** 212-215.

SUMMERFIELD, Q. 1987. *Some preliminaries to a comprehensive account of audio-visual speech perception,* 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.

SUMMERFIELD, Q. & MCGRATH, M. 1984. Detection and resolution of audio-visual incompatibility in the perception of vowels. *The Quarterly Journal of Experimental Psychology,* 36**,** 51-74.

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.

TEKI, S., CHAIT, M., KUMAR, S., VON KRIEGSTEIN, K. & GRIFFITHS, T. D. 2011. Brain Bases for Auditory Stimulus-Driven Figure-Ground Segregation. *Journal of Neuroscience,* 31**,** 164-171.

THEUNISSEN, F. E., DAVID, S. V., SINGH, N. C., HSU, A., VINJE, W. E. & GALLANT, J. L. 2001. Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. *Network: Computation in Neural Systems,* 12**,** 289-316.

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.

TIAN, X. & POEPPEL, D. 2012. Mental imagery of speech: linking motor and perceptual systems through internal simulation and estimation. *Frontiers in human neuroscience,* 6.

TIKHONOV, A. N. & ARSENIN, V. I. A. K. 1977. *Solutions of ill-posed problems,* Vh Winston.

TJAN, B. S., CHAO, E. & BERNSTEIN, L. E. 2014. A visual or tactile signal makes auditory speech detection more efficient by reducing uncertainty. *European Journal of Neuroscience,* 39**,** 1323-1331.

TOMITA, M. & EGGERMONT, J. J. 2005. Cross-correlation and joint spectro-temporal receptive field properties in auditory cortex. *Journal of neurophysiology,* 93**,** 378-392.

TYE-MURRAY, N., SOMMERS, M. & SPEHAR, B. 2007. Auditory and visual lexical neighborhoods in audiovisual speech perception. *Trends in Amplification,* 11**,** 233-241.

ULRICH, R., MILLER, J. & SCHRÖTER, H. 2007. Testing the race model inequality: an algorithm and computer programs. *Behavior Research Methods,* 39**,** 291-302.

VAN DER HORST, R., LEEUW, A. R. & DRESCHLER, W. A. 1999. Importance of temporal-envelope cues in consonant recognition. *The Journal of the Acoustical Society of America,* 105**,** 1801-1809.

VAN WASSENHOVE, V. 2013. Speech through ears and eyes: interfacing the senses with the supramodal brain. *Frontiers in psychology,* 4.

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.

VAN WASSENHOVE, V., GRANT, K. W. & POEPPEL, D. 2007. Temporal window of integration in auditory-visual speech perception. *Neuropsychologia,* 45**,** 598-607.

VATIKIOTIS-BATESON, E., EIGSTI, I.-M., YANO, S. & MUNHALL, K. G. 1998. Eye movement of perceivers during audiovisualspeech perception. *Perception & Psychophysics,* 60**,** 926-940.

WALDEN, B. E., PROSEK, R. A. & WORTHINGTON, D. W. 1975. Auditory and audiovisual feature transmission in hearing-impaired adults. *Journal of Speech, Language, and Hearing Research,* 18**,** 272-280.

WATKINS, K. E., STRAFELLA, A. P. & PAUS, T. 2003. Seeing and hearing speech excites the motor system involved in speech production. *Neuropsychologia,* 41**,** 989-994.

WELCH, R. B. & WARREN, D. H. 1980. Immediate perceptual response to intersensory discrepancy. *Psychol Bull,* 88**,** 638-667.

WINGATE, M., KIRBY, R. S., PETTYGROVE, S., CUNNIFF, C., SCHULZ, E., GHOSH, T., ROBINSON, C., LEE, L.-C., LANDA, R. & CONSTANTINO, J. 2014. Prevalence of Autism Spectrum Disorder Among Children Aged 8 Years-Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2010. *MMWR SURVEILLANCE SUMMARIES,* 63.

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,* 13**,** 1034-1043.

WU, M. C.-K., DAVID, S. V. & GALLANT, J. L. 2006. Complete functional characterization of sensory neurons by system identification. *Annu. Rev. Neurosci.,* 29**,** 477-505.

YANG, X., WANG, K. & SHAMMA, S. A. 1992. Auditory representations of acoustic signals. *Information Theory, IEEE Transactions on,* 38**,** 824-839.

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., SMILJANIC, R. & CHANDRASEKARAN, B. 2014. The neural processing of foreign-accented speech and its relationship to listener bias. *Frontiers in human neuroscience,* 8.

ZION-GOLUMBIC, E. M., COGAN, G. B., SCHROEDER, C. E. & POEPPEL, D. 2013a. Visual Input Enhances Selective Speech Envelope Tracking in Auditory Cortex at a "Cocktail Party". *Journal of Neuroscience,* 33**,** 1417-1426.

ZION-GOLUMBIC, E. M., DING, N., BICKEL, S., LAKATOS, P., SCHEVON, C. A., MCKHANN, G. M., GOODMAN, R. R., EMERSON, R., MEHTA, A. D., SIMON, J. Z., POEPPEL, D. & SCHROEDER, C. E. 2013b. Mechanisms Underlying Selective Neuronal Tracking of Attended Speech at a "Cocktail Party". *Neuron,* 77**,** 980-991.

Appendices

Appendix A

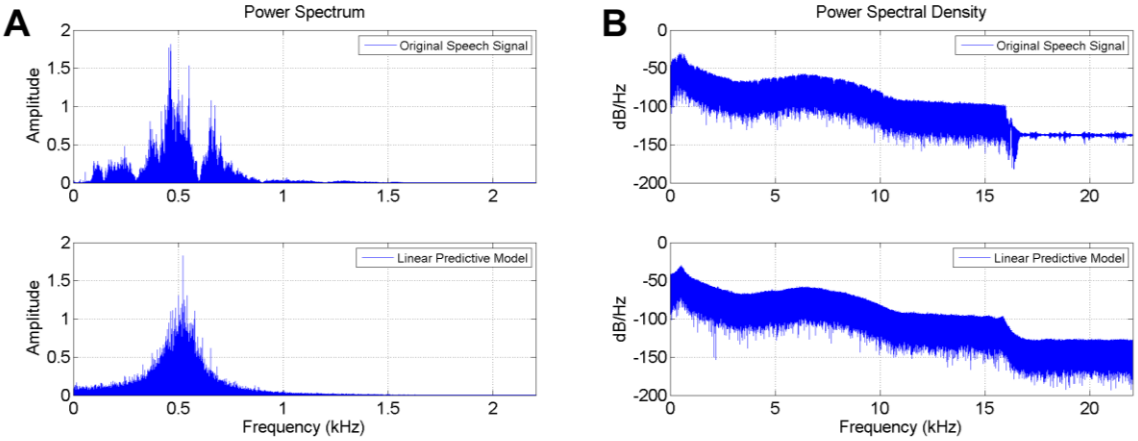


Figure A1: Spectral characteristics of the original speech signal and the linear predictive model.

A, Power spectra of the original (top) and modelled (bottom) signals. ***B***, Power spectral density of the original (top) and modelled (bottom) signals. The noise used to mask the speech signal was generated using a 50th order forward linear predictive model whose coefficients were estimated based on the original speech signal. This technique ensured that each frequency band was masked in an even manner. In other words, the spectrum of the noise was weighted according to the power in each frequency band in the speech signal. Hence, there is a clear correspondence between the spectral shape of the original and modelled signals.

Appendix B

```
function [model,t,c] = mTRFtrain(stim,resp,fs,map,tmin,tmax,lambda)
%mTRFtrain mTRF Toolbox training function.
%   MODEL = MTRFTRAIN(STIM,RESP,FS,MAP,TMIN,TMAX,LAMBDA) performs ridge
%   regression on the stimulus property STIM and the neural response data
%   RESP to solve for their linear mapping function MODEL. Pass in MAP==1
%   to map in the forward direction or MAP==-1 to map backwards. The
%   sampling frequency FS should be defined in Hertz and the time lags
%   should be set in milliseconds between TMIN and TMAX. Regularisation is
%   controlled by the ridge parameter LAMBDA.
%
%   [...,T,C] = MTRFTRAIN(...) also returns the vector of time lags T for
%   plotting MODEL and the regression constant C for absorbing any bias
%   when testing MODEL.
%
%   Inputs:
%   stim   - stimulus property (time by features)
%   resp   - neural response data (time by channels)
%   fs     - sampling frequency (Hz)
%   map    - mapping direction (forward==1, backward==-1)
%   tmin   - minimum time lag (ms)
%   tmax   - maximum time lag (ms)
%   lambda - ridge parameter
%
%   Outputs:
%   model  - linear mapping function (MAP==1: feats by lags by chans,
%            MAP==-1: chans by lags by feats)
%   t      - vector of time lags used (ms)
%   c      - regression constant
%
%   See README for examples of use.
%
%   See also LAGGEN MTRFPREDICT MTRFCROSSVAL MTRFMULTICROSSVAL.

%   References:
%      [1] Lalor EC, Pearlmutter BA, Reilly RB, McDarby G, Foxe JJ (2006).
%          The VESPA: a method for the rapid estimation of a visual evoked
%          potential. NeuroImage, 32:1549-1561.
%      [2] Lalor EC, Power AP, Reilly RB, Foxe JJ (2009). Resolving precise
%          temporal processing properties of the auditory system using
%          continuous stimuli. Journal of Neurophysiology, 102(1):349-359.

%   Author: Edmund Lalor, Michael Crosse, Giovanni Di Liberto
%   Lalor Lab, Trinity College Dublin, IRELAND
%   Email: edmundlalor@gmail.com
%   Website: http://lalorlab.net/
%   April 2014; Last revision: 08 January 2016

% Define x and y
if tmin > tmax
    error('Value of TMIN must be < TMAX')
end
if map == 1
    x = stim;
    y = resp;
elseif map == -1
    x = resp;
    y = stim;
    [tmin,tmax] = deal(tmax,tmin);
else
    error('Value of MAP must be 1 (forward) or -1 (backward)')
end
clear stim resp

% Convert time lags to samples
tmin = floor(tmin/1e3*fs*map);
tmax = ceil(tmax/1e3*fs*map);

% Generate lag matrix
X = [ones(size(x)),lagGen(x,tmin:tmax)];
```

```
% Set up regularisation
dim = size(X,2);
if size(x,2) == 1
    d = 2*eye(dim,dim);d([1,end]) = 1;
    u = [zeros(dim,1),eye(dim,dim-1)];
    l = [zeros(1,dim);eye(dim-1,dim)];
    M = d-u-l;
else
    M = eye(dim,dim);
end

% Calculate model
model = (X'*X+lambda*M)\(X'*y);

% Format outputs
c = model(1:size(x,2),:);
model = reshape(model(size(x,2)+1:end,:),size(x,2),length(tmin:tmax),size(y,2));
t = (tmin:tmax)/fs*1e3;

end
```

Appendix C

```
function [pred,r,p,mse] = mTRFpredict(stim,resp,model,fs,map,tmin,tmax,c)
%mTRFpredict mTRF Toolbox prediction function.
%   PRED = MTRFPREDICT(STIM,RESP,MODEL,FS,MAP,TMIN,TMAX,C) performs a
%   convolution of the stimulus property STIM or the neural response data
%   RESP with their linear mapping function MODEL to solve for the
%   prediction PRED. Pass in MAP==1 to predict RESP or MAP==-1 to predict
%   STIM. The sampling frequency FS should be defined in Hertz and the time
%   lags should be set in milliseconds between TMIN and TMAX. The
%   regression constant C absorbs any bias in MODEL.
%
%   [...,R,P,MSE] = MTRFPREDICT(...) also returns the correlation
%   coefficients R between the original and predicted values, the
%   corresponding p-values P and the mean squared errors MSE.
%
%   Inputs:
%   stim   - stimulus property (time by features)
%   resp   - neural response data (time by channels)
%   model  - linear mapping function (MAP==1: feats by lags by chans,
%            MAP==-1: chans by lags by feats)
%   fs     - sampling frequency (Hz)
%   map    - mapping direction (forward==1, backward==-1)
%   tmin   - minimum time lag (ms)
%   tmax   - maximum time lag (ms)
%   c      - regression constant
%
%   Outputs:
%   pred   - prediction (MAP==1: time by chans, MAP==-1: time by feats)
%   r      - correlation coefficients
%   p      - p-values of the correlations
%   mse    - mean squared errors
%
%   See README for examples of use.
%
%   See also LAGGEN MTRFTRAIN MTRFCROSSVAL MTRFMULTICROSSVAL.

%   Author: Michael Crosse, Giovanni Di Liberto
%   Lalor Lab, Trinity College Dublin, IRELAND
%   Email: edmundlalor@gmail.com
%   Website: http://lalorlab.net/
%   April 2014; Last revision: 08 January 2016

% Define x and y
if tmin > tmax
    error('Value of TMIN must be < TMAX')
end
if map == 1
    x = stim;
    y = resp;
elseif map == -1
    x = resp;
    y = stim;
    [tmin,tmax] = deal(tmax,tmin);
else
    error('Value of MAP must be 1 (forward) or -1 (backward)')
end

% Convert time lags to samples
tmin = floor(tmin/1e3*fs*map);
tmax = ceil(tmax/1e3*fs*map);

% Generate lag matrix
X = [ones(size(x)),lagGen(x,tmin:tmax)];

% Calculate prediction
model = [c;reshape(model,size(model,1)*size(model,2),size(model,3))];
pred = X*model;

% Calculate accuracy
if ~isempty(y)
```

```
    r = zeros(1,size(y,2));
    p = zeros(1,size(y,2));
    mse = zeros(1,size(y,2));
    for i = 1:size(y,2)
        [r(i),p(i)] = corr(y(:,i),pred(:,i));
        mse(i) = mean((y(:,i)-pred(:,i)).^2);
    end
end

end
```

Appendix D

```
function [r,p,mse,pred,model] = mTRFcrossval(stim,resp,fs,map,tmin,tmax,lambda)
%mTRFcrossval mTRF Toolbox cross-validation function.
%   [R,P,MSE] = MTRFCROSSVAL(STIM,RESP,FS,MAP,TMIN,TMAX,LAMBDA) performs
%   leave-one-out cross-validation on the set of stimuli STIM and the
%   neural responses RESP for the range of ridge parameter values LAMBDA.
%   As a measure of performance, it returns the correlation coefficients R
%   between the predicted and original signals, the corresponding p-values
%   P and the mean squared errors MSE. Pass in MAP==1 to map in the forward
%   direction or MAP==-1 to map backwards. The sampling frequency FS should
%   be defined in Hertz and the time lags should be set in milliseconds
%   between TMIN and TMAX.
%
%   [...,PRED,MODEL] = MTRFCROSSVAL(...) also returns the predictions PRED
%   and the linear mapping functions MODEL.
%
%   Inputs:
%   stim   - set of stimuli [cell{1,trials}(time by features)]
%   resp   - set of neural responses [cell{1,trials}(time by channels)]
%   fs     - sampling frequency (Hz)
%   map    - mapping direction (forward==1, backward==-1)
%   tmin   - minimum time lag (ms)
%   tmax   - maximum time lag (ms)
%   lambda - ridge parameter values
%
%   Outputs:
%   r      - correlation coefficients
%   p      - p-values of the correlations
%   mse    - mean squared errors
%   pred   - prediction [MAP==1: cell{1,trials}(lambdas by time by chans),
%            MAP==-1: cell{1,trials}(lambdas by time by feats)]
%   model  - linear mapping function (MAP==1: trials by lambdas by feats by
%            lags by chans, MAP==-1: trials by lambdas by chans by lags by
%            feats)
%
%   See README for examples of use.
%
%   See also LAGGEN MTRFTRAIN MTRFPREDICT MTRFMULTICROSSVAL.

%   Author: Michael Crosse
%   Lalor Lab, Trinity College Dublin, IRELAND
%   Email: edmundlalor@gmail.com
%   Website: http://lalorlab.net/
%   April 2014; Last revision: 31 May 2016

% Define x and y
if tmin > tmax
    error('Value of TMIN must be < TMAX')
end
if map == 1
    x = stim;
    y = resp;
elseif map == -1
    x = resp;
    y = stim;
    [tmin,tmax] = deal(tmax,tmin);
else
    error('Value of MAP must be 1 (forward) or -1 (backward)')
end
clear stim resp

% Convert time lags to samples
tmin = floor(tmin/1e3*fs*map);
tmax = ceil(tmax/1e3*fs*map);

% Set up regularisation
dim1 = size(x{1},2)*length(tmin:tmax)+size(x{1},2);
dim2 = size(y{1},2);
model = zeros(numel(x),numel(lambda),dim1,dim2);
if size(x{1},2) == 1
```

```
    d = 2*eye(dim1,dim1); d([1,end]) = 1;
    u = [zeros(dim1,1),eye(dim1,dim1-1)];
    l = [zeros(1,dim1);eye(dim1-1,dim1)];
    M = d-u-l;
else
    M = eye(dim1,dim1);
end

% Training
X = cell(1,numel(x));
for i = 1:numel(x)
    % Generate lag matrix
    X{i} = [ones(size(x{i})),lagGen(x{i},tmin:tmax)];
    % Calculate model for each lambda value
    for j = 1:length(lambda)
        model(i,j,:,:) = (X{i}'*X{i}+lambda(j)*M)\(X{i}'*y{i});
    end
end

% Testing
pred = cell(1,numel(x));
r = zeros(numel(x),numel(lambda),dim2);
p = zeros(numel(x),numel(lambda),dim2);
mse = zeros(numel(x),numel(lambda),dim2);
for i = 1:numel(x)
    pred{i} = zeros(numel(lambda),size(y{i},1),dim2);
    % Define training trials
    trials = 1:numel(x);
    trials(i) = [];
    % Perform cross-validation for each lambda value
    for j = 1:numel(lambda)
        % Calculate prediction
        pred{i}(j,:,:) = X{i}*squeeze(mean(model(trials,j,:,:)));
        % Calculate accuracy
        for k = 1:dim2
            [r(i,j,k),p(i,j,k)] = corr(y{i}(:,k),squeeze(pred{i}(j,:,k))');
            mse(i,j,k) = mean((y{i}(:,k)-squeeze(pred{i}(j,:,k))').^2);
        end
    end
end

end
```

Appendix E

```
function [r,p,mse,pred,model] =
mTRFmulticrossval(stim,resp,resp1,resp2,fs,tmin,tmax,lambda1,lambda2)
%mTRFmulticrossval mTRF Toolbox multisensory cross-validation function.
%   [R,P,MSE] = MTRFMULTICROSSVAL(STIM,RESP,RESP1,RESP2,FS,MAP,TMIN,TMAX,
%   LAMBDA1,LAMBDA2) performs leave-one-out cross-validation of an
%   additive model for a multisensory dataset as follows:
%   1. Separate unisensory models are calculated using the set of stimuli
%      STIM and unisensory neural responses RESP1 and RESP2 for the range
%      of ridge parameter values LAMBDA1 and LAMBDA2 respectively.
%   2. The algebraic sums of the unisensory models for every combination of
%      LAMBDA1 and LAMBDA2 are calculated, i.e., the additive models.
%   3. The additive models are validated by testing them on the set of
%      multisensory neural responses RESP.
%   As a measure of performance, it returns the correlation coefficients R
%   between the predicted and original signals, the corresponding p-values
%   P and the mean squared errors MSE. The time lags T should be set in
%   milliseconds between TMIN and TMAX and the sampling frequency FS should
%   be defined in Hertz. Pass in MAP==1 to map in the forward direction or
%   MAP==-1 to map backwards. The neural responses in all three sensory
%   conditions must have been recorded for the same set of stimuli STIM.
%
%   [...,PRED,MODEL] = MTRFMULTICROSSVAL(...) also returns the predictions
%   PRED and the linear mapping functions MODEL.
%
%   Inputs:
%   stim   - set of stimuli [cell{1,trials}(time by features)]
%   resp   - set of multisensory neural responses [cell{1,trials}(time by channels)]
%   resp1  - set of unisensory 1 neural responses [cell{1,trials}(time by channels)]
%   resp2  - set of unisensory 2 neural responses [cell{1,trials}(time by channels)]
%   fs     - sampling frequency (Hz)
%   map    - mapping direction (forward==1, backward==-1)
%   tmin   - minimum time lag (ms)
%   tmax   - maximum time lag (ms)
%   lambda1- unisensory 1 ridge parameter values
%   lambda2- unisensory 2 ridge parameter values
%
%   Outputs:
%   r      - correlation coefficients
%   p      - p-values of the correlations
%   mse    - mean squared errors
%   pred   - prediction [MAP==1: cell{1,trials}(lambdas1 by lambdas2 by
%            time by chans), MAP==-1: cell{1,trials}(lambdas1 by lambdas2
%            by time by feats)]
%   model  - linear mapping function (MAP==1: trials by lambdas1 by
%            lambdas2 by feats by lags by chans, MAP==-1: trials by
%            lambdas1 by lambdas2 by chans by lags by feats)
%
%   See README for examples of use.
%
%   See also LAGGEN MTRFTRAIN MTRFPREDICT MTRFCROSSVAL.

%   Author: Michael Crosse
%   Lalor Lab, Trinity College Dublin, IRELAND
%   Email: edmundlalor@gmail.com
%   Website: http://lalorlab.net/
%   April 2014; Last revision: 01 June 2016

% Define x and y
if tmin > tmax
    error('Value of TMIN must be < TMAX')
end
if map == 1
    x = stim;
    y = resp;
elseif map == -1
    x = resp;
    y = stim;
    [tmin,tmax] = deal(tmax,tmin);
else
```

```
    error('Value of MAP must be 1 (forward) or -1 (backward)')
end
clear stim resp

% Convert time lags to samples
tmin = floor(tmin/1e3*fs*map);
tmax = ceil(tmax/1e3*fs*map);

% Set up regularisation
dim1 = size(x{1},2)*length(tmin:tmax)+size(x{1},2);
dim2 = size(y{1},2);
model = zeros(numel(x),numel(lambda1),numel(lambda2),dim1,dim2);
if size(x{1},2) == 1
    d = 2*eye(dim1,dim1); d([1,end]) = 1;
    u = [zeros(dim1,1),eye(dim1,dim1-1)];
    l = [zeros(1,dim1);eye(dim1-1,dim1)];
    M = d-u-l;
else
    M = eye(dim1,dim1);
end

% Training
X = cell(1,numel(x));
for i = 1:numel(x)
    % Generate lag matrix
    X{i} = [ones(size(x{i})),lagGen(x{i},tmin:tmax)];
    if map == 1
        % Calculate unisensory models for each lambda value
        model1 = zeros(numel(lambda1),dim1,dim2);
        for j = 1:numel(lambda1)
            model1(j,:,:) = (X'*X+lambda1(j)*M)\X'*resp1{i};
        end
        model2 = zeros(numel(lambda2),dim1,dim2);
        for j = 1:numel(lambda2)
            model2(j,:,:) = (X'*X+lambda2(j)*M)\X'*resp2{i};
        end
    elseif map == -1
        % Generate lag matrices
        X1 = [ones(size(resp1{i})),lagGen(resp1{i},tmin:tmax)];
        X2 = [ones(size(resp2{i})),lagGen(resp2{i},tmin:tmax)];
        % Calculate unisensory models for each lambda value
        model1 = zeros(numel(lambda1),dim1,dim2);
        for j = 1:numel(lambda1)
            model1(j,:,:) = (X1'*X1+lambda1(j)*M)\X1'*y{i};
        end
        model2 = zeros(numel(lambda2),dim1,dim2);
        for j = 1:numel(lambda2)
            model2(j,:,:) = (X2'*X2+lambda2(j)*M)\X2'*y{i};
        end
        clear X1 X2
    end
    % Sum unisensory models for every combination of lambda values
    for j = 1:numel(lambda1)
        for k = 1:numel(lambda2)
            model(i,j,k,:,:) = model1(j,:,:)+model2(k,:,:);
        end
    end
    clear model1 model2
end
clear resp1 resp2

% Testing
pred = cell(1,numel(x));
r = zeros(numel(x),numel(lambda1),numel(lambda2),dim2);
p = zeros(numel(x),numel(lambda1),numel(lambda2),dim2);
mse = zeros(numel(x),numel(lambda1),numel(lambda2),dim2);
for i = 1:numel(x)
    pred{i} = zeros(numel(lambda1),numel(lambda2),size(y{i},1),dim2);
    % Define training trials
    trials = 1:numel(x);
    trials(i) = [];
    % Perform cross-validation for every combination of lambda values
    for j = 1:numel(lambda1)
        for k = 1:numel(lambda2)
            % Calculate prediction
            pred{i}(j,k,:,:) = X{i}*squeeze(mean(model(trials,j,k,:,:)));
            % Calculate accuracy
            for l = 1:dim2
                [r(i,j,k,l),p(i,j,k,l)] = corr(y{i}(:,l),squeeze(pred{i}(j,k,:,l))');
                mse(i,j,k,l) = mean((y{i}(:,l)-squeeze(pred{i}(j,k,:,l))').^2);
            end
        end
    end
end

end
```

Appendix F

```
function xLag = lagGen(x,lags)
%lagGen Lag generator.
%   [XLAG] = LAGGEN(X,LAGS) returns the matrix XLAG containing the lagged
%   time series of X for a range of time lags given by the vector LAGS. If
%   X is multivariate, LAGGEN will concatenate the features for each lag
%   along the columns of XLAG.
%
%   Inputs:
%   x    - vector or matrix of time series data (time by features)
%   lags - vector of integer time lags (samples)
%
%   Outputs:
%   xLag - matrix of lagged time series data (time by lags*feats)
%
%   See README for examples of use.
%
%   See also MTRFTRAIN MTRFPREDICT MTRFCROSSVAL MTRFMULTICROSSVAL.

%   Author: Michael Crosse
%   Lalor Lab, Trinity College Dublin, IRELAND
%   Email: edmundlalor@gmail.com
%   Website:

://lalorlab.net/
%   April 2014; Last revision: 18 August 2015

xLag = zeros(size(x,1),size(x,2)*length(lags));

i = 1;
for j = 1:length(lags)
    if lags(j) < 0
        xLag(1:end+lags(j),i:i+size(x,2)-1) = x(-lags(j)+1:end,:);
    elseif lags(j) > 0
        xLag(lags(j)+1:end,i:i+size(x,2)-1) = x(1:end-lags(j),:);
    else
        xLag(:,i:i+size(x,2)-1) = x;
    end
    i = i+size(x,2);
end

end
```

Appendix G

The effects of congruency on the latency of continuous audiovisual speech processing

Michael J Crosse[1,2], Edmund C Lalor[1,2,3]

[1] School of Engineering, [2] Trinity Centre for Bioengineering and [3] Trinity College Institute of Neuroscience, Trinity College Dublin, Dublin 2, Ireland.

Introduction

- EEG research has shown that the presentation of congruent audiovisual (AVc) speech can facilitate an earlier cortical response compared to that of audio-only (A) speech [1, 2].
- A similar latency effect has been shown in event-averaged EEG measures calculated in response to incongruent audiovisual (AVi) syllables [3].

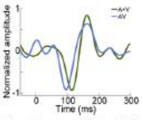


Fig: Crosse and Lalor [2]

- The temporal coherence between the bimodal streams may serve to induce crossmodal interactions despite the conflicting information and allow for earlier processing of the speech information [4].
- Here we build on previous work which has looked at semantically incongruent multisensory pairings [5], and examine also effects of temporal incongruence on the latency of auditory processing using a response measure suited to continuous and natural speech [6].

Methods

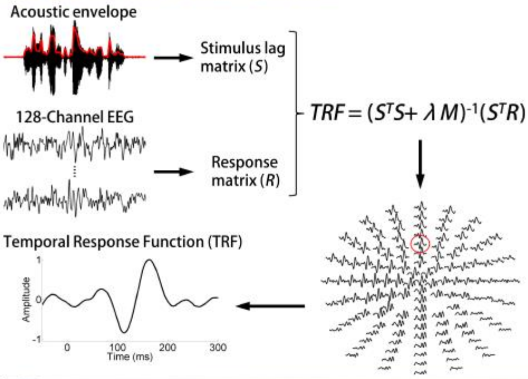
Experimental Procedure

- 128-channel EEG was recorded from 14 subjects (5 female, aged 21—37 years).
- 15 minutes of natural speech was presented in 7 different conditions, each with differing levels of AV congruency.

Audiovisual Stimuli

CONDITION	AUDIO	VIDEO	CONGRUENCY
Audiovisual congruent (AVc)	Obama audio *n*		
Audiovisual incongruent (AVi)	Obama audio *n*		
Audiovisual incongruent female (AVif)	Obama audio *n*		
Audiovisual incongruent nature (AVin)	Obama audio *n*		
Audiovisual static face (AVsf)	Obama audio *n*		

Temporal Response Function Estimation



Acoustic envelope

Stimulus lag matrix (*S*)

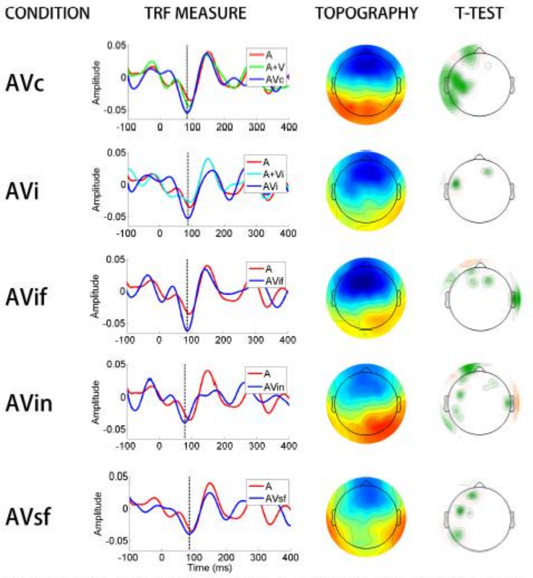
128-Channel EEG

Response matrix (*R*)

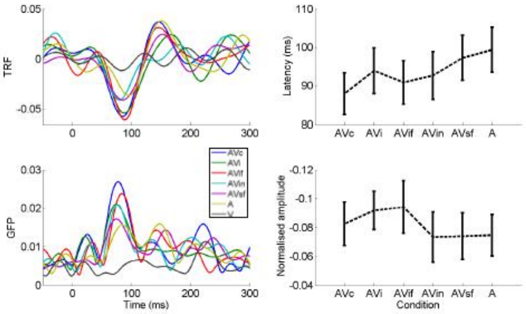
$TRF = (S^TS + \lambda M)^{-1}(S^TR)$

Temporal Response Function (TRF)

Results



TRFs for each AV condition versus A condition. Topographies show scalp distribution at ~85 ms (dotted line), scaled to TRF axes. Paired *t*-tests of AV−A amplitude from 50 to 100ms plotted on scalp maps.



Butterfly plots of TRFs and GFPs. Peak latencies and amplitudes of negative TRF component for each condition. Error bars represent SEM.

Discussion

- Semantic congruency of AV pairings appears to modulate the latency of the auditory response, although further investigation is required.
- Temporal coherence between the bimodal streams is not necessary to facilitate response latency suggesting an alternative mechanism.
- Supra-additive response measures are only evident in conditions that include dynamic human faces, congruent or incongruent.

References

[1] van Wassenhove V, Grant KW, Poeppel D (2005) *PNAS* 103(4):1181—1186.
[2] Crosse MJ, Lalor EC (2014) *J Neurophysiol* 11:1400—1408.
[3] Stekelenburg JJ, Vroomen J (2007) *J Cog Neurosci* 19(12):1964—1973.
[4] Shamma SA, Elhilali M, Micheyl C (2011) *Trends Neurosci* 34(3):114—123.
[5] Fiebelkorn IC, Foxe JJ, Molholm S (2010) *Cereb Cortex* 20(1):109—20.
[6] Lalor EC, Foxe JJ (2010) *Eur J Neurosci* 31(1):189—193.

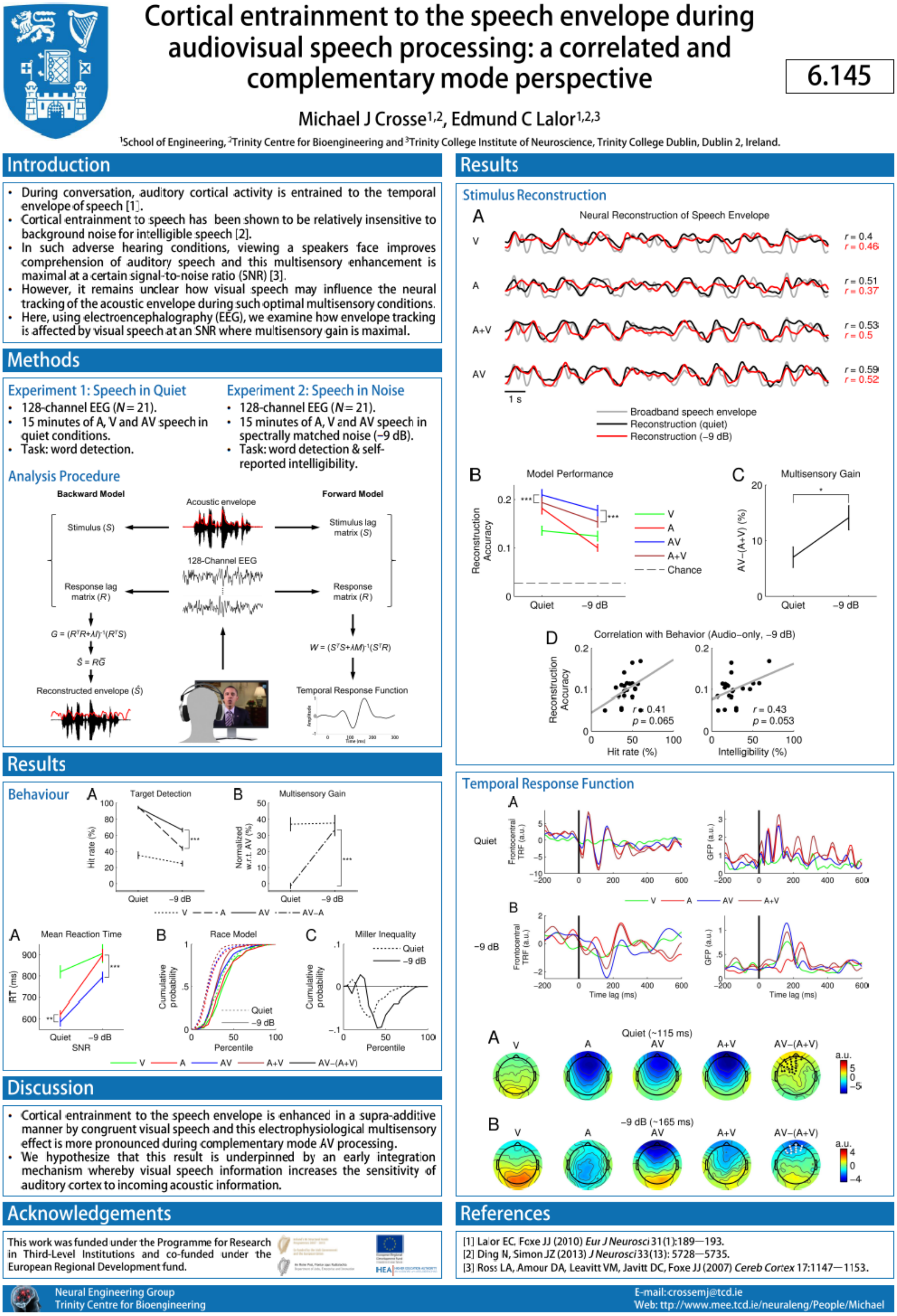
Acknowledgements

This work was funded under the Programme for Research in Third-Level Institutions and co-funded under the European Regional Development fund.

Neural Engineering Group
Trinity Centre for Bioengineering

E-mail: crossemj@tcd.ie
Web: http://www.mee.tcd.ie/neuraleng/People/Michael

Appendix H



Cortical entrainment to the speech envelope during audiovisual speech processing: a correlated and complementary mode perspective

6.145

Michael J Crosse[1,2], Edmund C Lalor[1,2,3]

[1]School of Engineering, [2]Trinity Centre for Bioengineering and [3]Trinity College Institute of Neuroscience, Trinity College Dublin, Dublin 2, Ireland.

Introduction

- During conversation, auditory cortical activity is entrained to the temporal envelope of speech [1].
- Cortical entrainment to speech has been shown to be relatively insensitive to background noise for intelligible speech [2].
- In such adverse hearing conditions, viewing a speakers face improves comprehension of auditory speech and this multisensory enhancement is maximal at a certain signal-to-noise ratio (SNR) [3].
- However, it remains unclear how visual speech may influence the neural tracking of the acoustic envelope during such optimal multisensory conditions.
- Here, using electroencephalography (EEG), we examine how envelope tracking is affected by visual speech at an SNR where multisensory gain is maximal.

Methods

Experiment 1: Speech in Quiet

- 128-channel EEG ($N = 21$).
- 15 minutes of A, V and AV speech in quiet conditions.
- Task: word detection.

Experiment 2: Speech in Noise

- 128-channel EEG ($N = 21$).
- 15 minutes of A, V and AV speech in spectrally matched noise (−9 dB).
- Task: word detection & self-reported intelligibility.

Analysis Procedure

Results

Behaviour

Discussion

- Cortical entrainment to the speech envelope is enhanced in a supra-additive manner by congruent visual speech and this electrophysiological multisensory effect is more pronounced during complementary mode AV processing.
- We hypothesize that this result is underpinned by an early integration mechanism whereby visual speech information increases the sensitivity of auditory cortex to incoming acoustic information.

Acknowledgements

This work was funded under the Programme for Research in Third-Level Institutions and co-funded under the European Regional Development fund.

Results

Stimulus Reconstruction

Temporal Response Function

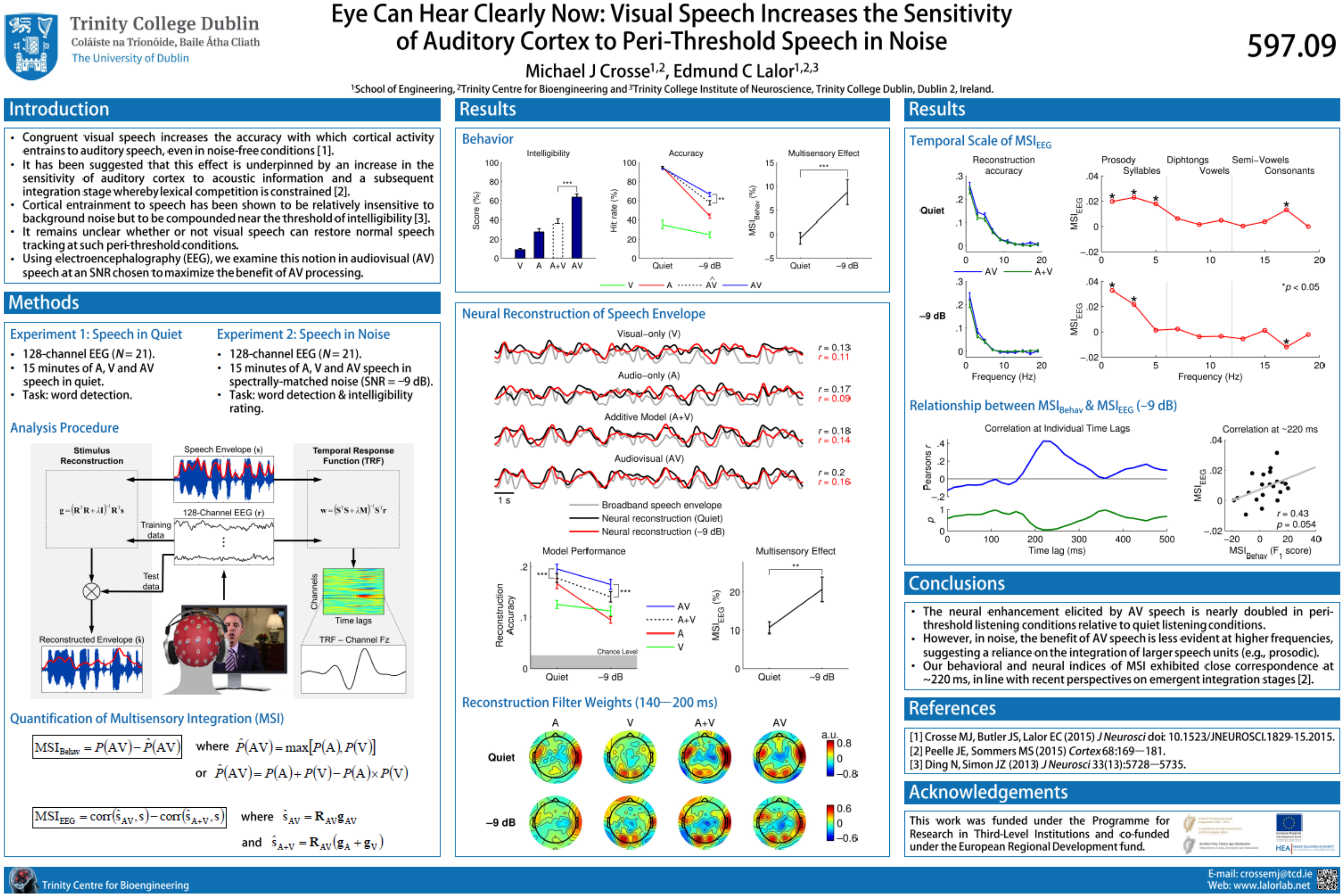
References

[1] Lalor EC, Foxe JJ (2010) *Eur J Neurosci* 31(1):189—193.
[2] Ding N, Simon JZ (2013) *J Neurosci* 33(13): 5728—5735.
[3] Ross LA, Amour DA, Leavitt VM, Javitt DC, Foxe JJ (2007) *Cereb Cortex* 17:1147—1153.

Neural Engineering Group
Trinity Centre for Bioengineering

E-mail: crossemj@tcd.ie
Web: ttp://www.mee.tcd.ie/neuraleng/People/Michael

Appendix I



Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Eye Can Hear Clearly Now: Visual Speech Increases the Sensitivity of Auditory Cortex to Peri-Threshold Speech in Noise

597.09

Michael J Crosse[1,2], Edmund C Lalor[1,2,3]

[1]School of Engineering, [2]Trinity Centre for Bioengineering and [3]Trinity College Institute of Neuroscience, Trinity College Dublin, Dublin 2, Ireland.

Introduction

- Congruent visual speech increases the accuracy with which cortical activity entrains to auditory speech, even in noise-free conditions [1].
- It has been suggested that this effect is underpinned by an increase in the sensitivity of auditory cortex to acoustic information and a subsequent integration stage whereby lexical competition is constrained [2].
- Cortical entrainment to speech has been shown to be relatively insensitive to background noise but to be compounded near the threshold of intelligibility [3].
- It remains unclear whether or not visual speech can restore normal speech tracking at such peri-threshold conditions.
- Using electroencephalography (EEG), we examine this notion in audiovisual (AV) speech at an SNR chosen to maximize the benefit of AV processing.

Methods

Experiment 1: Speech in Quiet

- 128-channel EEG (*N* = 21).
- 15 minutes of A, V and AV speech in quiet.
- Task: word detection.

Experiment 2: Speech in Noise

- 128-channel EEG (*N* = 21).
- 15 minutes of A, V and AV speech in spectrally-matched noise (SNR = −9 dB).
- Task: word detection & intelligibility rating.

Analysis Procedure

Stimulus Reconstruction
$\mathbf{g} = (\mathbf{R}^T\mathbf{R} + \lambda\mathbf{I})^{-1}\mathbf{R}^T\mathbf{s}$
Speech Envelope (s)
128-Channel EEG (r)
Training data
Test data
Temporal Response Function (TRF)
$\mathbf{w} = (\mathbf{S}^T\mathbf{S} + \lambda\mathbf{M})^{-1}\mathbf{S}^T\mathbf{r}$
Channels
Time lags
TRF – Channel Fz
Reconstructed Envelope (ŝ)

Quantification of Multisensory Integration (MSI)

$$\boxed{\mathrm{MSI_{Behav}} = P(\mathrm{AV}) - \hat{P}(\mathrm{AV})}$$ where $\hat{P}(\mathrm{AV}) = \max[P(\mathrm{A}), P(\mathrm{V})]$

or $\hat{P}(\mathrm{AV}) = P(\mathrm{A}) + P(\mathrm{V}) - P(\mathrm{A}) \times P(\mathrm{V})$

$$\boxed{\mathrm{MSI_{EEG}} = \mathrm{corr}(\hat{s}_{\mathrm{AV}}, s) - \mathrm{corr}(\hat{s}_{\mathrm{A+V}}, s)}$$ where $\hat{s}_{\mathrm{AV}} = \mathbf{R}_{\mathrm{AV}}\mathbf{g}_{\mathrm{AV}}$

and $\hat{s}_{\mathrm{A+V}} = \mathbf{R}_{\mathrm{AV}}(\mathbf{g}_{\mathrm{A}} + \mathbf{g}_{\mathrm{V}})$

Results

Behavior

Neural Reconstruction of Speech Envelope

Reconstruction Filter Weights (140—200 ms)

Results

Temporal Scale of $\mathrm{MSI_{EEG}}$

Relationship between $\mathrm{MSI_{Behav}}$ & $\mathrm{MSI_{EEG}}$ (−9 dB)

Conclusions

- The neural enhancement elicited by AV speech is nearly doubled in peri-threshold listening conditions relative to quiet listening conditions.
- However, in noise, the benefit of AV speech is less evident at higher frequencies, suggesting a reliance on the integration of larger speech units (e.g., prosodic).
- Our behavioral and neural indices of MSI exhibited close correspondence at ~220 ms, in line with recent perspectives on emergent integration stages [2].

References

[1] Crosse MJ, Butler JS, Lalor EC (2015) *J Neurosci* doi: 10.1523/JNEUROSCI.1829-15.2015.
[2] Peelle JE, Sommers MS (2015) *Cortex* 68:169—181.
[3] Ding N, Simon JZ (2013) *J Neurosci* 33(13):5728—5735.

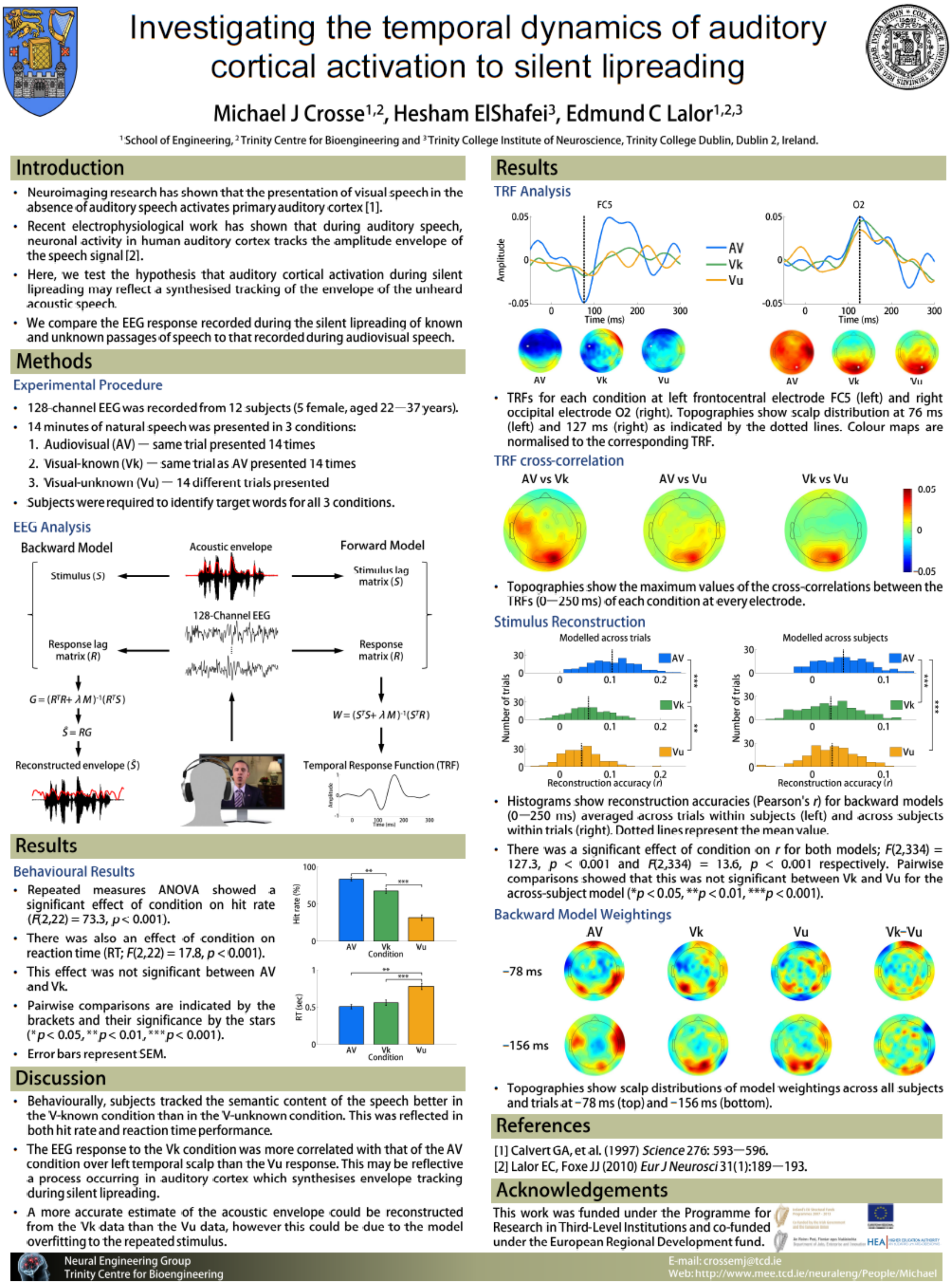
Acknowledgements

This work was funded under the Programme for Research in Third-Level Institutions and co-funded under the European Regional Development fund.

Trinity Centre for Bioengineering

E-mail: crossemj@tcd.ie
Web: www.lalorlab.net

Appendix J



Investigating the temporal dynamics of auditory cortical activation to silent lipreading

Michael J Crosse[1,2], Hesham ElShafei[3], Edmund C Lalor[1,2,3]

[1] School of Engineering, [2] Trinity Centre for Bioengineering and [3] Trinity College Institute of Neuroscience, Trinity College Dublin, Dublin 2, Ireland.

Introduction

- Neuroimaging research has shown that the presentation of visual speech in the absence of auditory speech activates primary auditory cortex [1].
- Recent electrophysiological work has shown that during auditory speech, neuronal activity in human auditory cortex tracks the amplitude envelope of the speech signal [2].
- Here, we test the hypothesis that auditory cortical activation during silent lipreading may reflect a synthesised tracking of the envelope of the unheard acoustic speech.
- We compare the EEG response recorded during the silent lipreading of known and unknown passages of speech to that recorded during audiovisual speech.

Methods

Experimental Procedure

- 128-channel EEG was recorded from 12 subjects (5 female, aged 22—37 years).
- 14 minutes of natural speech was presented in 3 conditions:
  1. Audiovisual (AV) — same trial presented 14 times
  2. Visual-known (Vk) — same trial as AV presented 14 times
  3. Visual-unknown (Vu) — 14 different trials presented
- Subjects were required to identify target words for all 3 conditions.

EEG Analysis

Backward Model

Acoustic envelope

Forward Model

Stimulus (S)

Stimulus lag matrix (S)

128-Channel EEG

Response lag matrix (R)

Response matrix (R)

$G = (R^TR + \lambda M)^{-1}(R^TS)$

$\hat{S} = RG$

$W = (S^TS + \lambda M)^{-1}(S^TR)$

Reconstructed envelope ($\hat{S}$)

Temporal Response Function (TRF)

Results

Behavioural Results

- Repeated measures ANOVA showed a significant effect of condition on hit rate ($F(2,22) = 73.3$, $p < 0.001$).
- There was also an effect of condition on reaction time (RT; $F(2,22) = 17.8$, $p < 0.001$).
- This effect was not significant between AV and Vk.
- Pairwise comparisons are indicated by the brackets and their significance by the stars (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$).
- Error bars represent SEM.

Discussion

- Behaviourally, subjects tracked the semantic content of the speech better in the V-known condition than in the V-unknown condition. This was reflected in both hit rate and reaction time performance.
- The EEG response to the Vk condition was more correlated with that of the AV condition over left temporal scalp than the Vu response. This may be reflective a process occurring in auditory cortex which synthesises envelope tracking during silent lipreading.
- A more accurate estimate of the acoustic envelope could be reconstructed from the Vk data than the Vu data, however this could be due to the model overfitting to the repeated stimulus.

Results

TRF Analysis

- TRFs for each condition at left frontocentral electrode FC5 (left) and right occipital electrode O2 (right). Topographies show scalp distribution at 76 ms (left) and 127 ms (right) as indicated by the dotted lines. Colour maps are normalised to the corresponding TRF.

TRF cross-correlation

- Topographies show the maximum values of the cross-correlations between the TRFs (0—250 ms) of each condition at every electrode.

Stimulus Reconstruction

- Histograms show reconstruction accuracies (Pearson's r) for backward models (0—250 ms) averaged across trials within subjects (left) and across subjects within trials (right). Dotted lines represent the mean value.
- There was a significant effect of condition on r for both models; $F(2,334) = 127.3$, $p < 0.001$ and $F(2,334) = 13.6$, $p < 0.001$ respectively. Pairwise comparisons showed that this was not significant between Vk and Vu for the across-subject model (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$).

Backward Model Weightings

- Topographies show scalp distributions of model weightings across all subjects and trials at −78 ms (top) and −156 ms (bottom).

References

[1] Calvert GA, et al. (1997) *Science* 276: 593—596.

[2] Lalor EC, Foxe JJ (2010) *Eur J Neurosci* 31(1):189—193.

Acknowledgements

This work was funded under the Programme for Research in Third-Level Institutions and co-funded under the European Regional Development fund.

Neural Engineering Group
Trinity Centre for Bioengineering

E-mail: crossemj@tcd.ie
Web: http://www.mee.tcd.ie/neuraleng/People/Michael