Trinity College Dublin

The University of Dublin

Investigating the Neural Correlates of Speech Processing and Selective Auditory Attention using Electroencephalography (EEG)

Emily S. Teoh, B.E., M.Sc.

Under the supervision of Dr. Edmund C. Lalor

A thesis presented to the

University of Dublin

In fulfilment of the requirements for the degree of

Doctor of Philosophy

2020

Department of Electronic and Electrical Engineering

University of Dublin, Trinity College

Declaration

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

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

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

Signed,

____

Emily Teoh

16th March 2020

Summary

Speech comprehension is a remarkable human ability. Most normal-hearing people are adept at attending to a speech stream even amidst a noisy multi-talker background and parsing the layers of information it contains in real-time to uncover its meaning. The cortex is thought to be hierarchically organised to perform the latter feat, with higher levels processing increasingly abstract features. Nonetheless, the precise computations it undertakes and how these processes are modulated by attention remain incompletely understood.

The development of novel encoding/decoding approaches for relating high temporal resolution neuroimaging modalities to representations of multivariate stimuli like continuous natural speech have enabled researchers to shed some light on this topic. An important recent discovery is that low-frequency, non-invasive EEG/MEG tracks the amplitude envelope of speech (a low-level acoustic measure conveying many cues important for speech comprehension) and that a robust reconstruction of the measure can be acquired from neural data. Moreover, during cocktail party listening tasks, this tracking and the accuracy of reconstruction have been shown to be modulated by attention. These findings have given rise to further research along several lines, including studies employing the framework to investigate how cortex encodes other speech processing stages along the hierarchy and how this encoding is affected by attention, as well as studies developing the idea of exploiting envelope reconstruction as a means of decoding attentional selection for the implementation of ‘smart’ devices like steerable hearing aids.

In this thesis, we employ a particular encoding/decoding approach – the temporal response function (TRF) – in conjunction with EEG to address several questions within these subareas. We first tested the efficacy of a state-of-the-art framework for utilising envelope reconstruction to decode auditory attention (O’Sullivan et al., 2015) within the context of a cocktail party paradigm with moving talkers and showed that it is robust. This was motivated by the non-stationarity of real-world environments. We then considered if (1) the decoder weights themselves (i.e., the model weights mapping from EEG data to the acoustic envelope) and (2) alpha power might contain unique information that can be leveraged for the decoding of attention. We showed that

incorporating a metric based on the consistency of model weights across subjects into the decoding framework yielded an improvement in performance above and beyond using envelope reconstruction alone.

We then investigated the neural processing of prosody – an aspect of spoken language that conveys another layer of meaning on top of linguistic units. In particular, we were interested to explore how prosodic pitch is encoded in low-frequency EEG during listening to continuous natural speech. We mapped two measures of prosodic pitch – relative pitch and harmonic resolvability – to concurrently-recorded EEG. These measures were inspired by an ECoG study showing neural tracking of relative pitch during listening to sentences (Tang, Hamilton, & Chang, 2017), and an fMRI study demonstrating that there are cortical regions that respond primarily to resolved harmonics (Norman-Haignere, Kanwisher, & McDermott, 2013). We found that delta-band EEG tracks relative pitch during listening to continuous natural speech, and that this tracking is dissociable from the tracking of other acoustic and phonetic features.

Finally, we tested how attention modulates EEG signatures hypothesized to represent processing at the pre-lexical level, as well as the signature of relative pitch found in the previous study. The former inquiry was motivated by the longstanding debates as to whether attention modulation operates at an early or late stage in the speech processing hierarchy, and whether an acoustic-phonetic transformation stage occurs as an intermediate step in mapping the acoustic speech signal to the mental store of words. We found that a phonetic feature representation could uniquely predict neural activity above and beyond other measures. Additionally, this unique predictive power was significantly modulated by attention, unlike that of a representation derived from acoustics that had been suggested to explain away the contribution of phonetic features (Daube et al., 2019). This lends support to the notion of the phonetic feature representation being a distinct and higher-level stage in the hierarchy. We also found attentional modulation of our signature of relative pitch and showed that incorporating the reconstruction accuracy of this representation into the decoding attentional selection framework led to a small improvement in decoding performance for some subjects.

Acknowledgements

This thesis would not have been possible without the invaluable support and contributions of many people. I am grateful to:

Ed, thank you for your constant guidance, trust, encouragement, and infectious enthusiasm. It has been five really enjoyable years, and I have learned so much from you.

Richard, thank you for all your kind help, especially with the move.

Gio, thank you for patiently answering all my questions at the start of the PhD.

Lalor Lab members past and present: Mick, James, Ger, Denis, Adam, Aisling, Andy, Kevin, Nate, Aaron, Shy, Lauren, Michael, for all the helpful feedback, comments, discussions, and for making the lab such a nice place to work.

Shruti, for friendship, the many times you've put me up, and all the fun times in the lab and in so many other places.

Reilly Lab members past and present, for making the Dublin lab a great place to work.

The volunteers and subjects in these studies, for giving up your time and sitting through the long experiments.

The Teos, thank you for making Dublin feel like another home.

Finally, Mum, Dad, and Karen, thank you for your love and support, and for inspiring me and believing in me.

This research was supported by the Irish Research Council (GOIPG/2015/3378).

Emily Teoh

Sept 2019

Publications Arising from this Thesis

Journal Articles

- **Teoh, E. S.**, Cappelloni, M. S., & Lalor, E. C. (2019). Prosodic pitch processing is represented in delta-band EEG and is dissociable from the cortical tracking of other acoustic and phonetic features. *European Journal of Neuroscience*, in press. (Chapter 4)
- **Teoh, E. S.**, & Lalor, E. C. (2019). EEG decoding of the target speaker in a cocktail party scenario: Considerations regarding dynamic switching of talker location. *Journal of Neural Engineering*. (Chapter 3)
- Zuk, N. J., **Teoh, E. S.**, & Lalor, E. C. (2020). EEG-based classification of natural sounds reveals specialized responses to speech and music. *Neuroimage.*

In Preparation/Review

- **Teoh, E. S.**, & Lalor, E. C. (in prep.). Attentional effects dissociate the acoustic and phonetic processing stages of speech comprehension. (Chapter 5)

Conference Poster Presentations

- **Teoh, E. S.**, & Lalor, E. C. Leveraging the dissociability of pitch and intensity in cortex to improve the decoding of cocktail party attention. Society for Neuroscience 46th annual meeting, San Diego, USA, November 2016
- **Teoh, E. S.**, & Lalor, E. C. Effects of Selective Auditory Attention on Phoneme-Level Processing in a Multi-speaker Environment. Auditory Cortex, Banff, Canada, September 2017
- **Teoh, E. S.**, & Lalor, E. C. Effects of Selective Auditory Attention on Phoneme-Level Processing in a Multi-speaker Environment. Advances and Perspectives in Auditory Neuroscience, Washington DC, USA, November 2017

- **Teoh, E. S.**, & Lalor, E. C. Selective Auditory Attention in a Dynamic, Multispeaker Environment. Society for Neuroscience 47th annual meeting, Washingon DC, USA, November 2017

Contents

List of Figures

Glossary of Acronyms

A1	Primary auditory cortex
AEP	Auditory Evoked Potential
ATL	Anterior Temporal Lobe
ECoG	Electrocorticography
EEG	Electroencephalography
env	Envelope
ERP/F	Event-related potential/field
f	Phonetic features
fMRI	Functional Magnetic Resonance Imaging
fo	Phoneme onset
$\boldsymbol{f_{rel}}$	Relative pitch
HG	Heschl’s Gyrus
HRTF	Head-related transfer function
IFG	Inferior Frontal Gyrus
ITG	Inferior Temporal Gyrus
IPA	International Phonetic Alphabet
LTI	Linear time-invariant
MEG	Magnetoencephalography
MGN	Medial Geniculate Nucleus
MSE	Mean squared error
MTG	Medial Temporal Gyrus
s	Spectrogram
sD	Spectrogram derivative
SNR	Signal-to-noise ratio
STG	Superior Temporal Gyrus
STS	Superior Temporal Sulcus
SVM	Support Vector Machine
TRF	Temporal response function

Chapter 1 Introduction

Speech is incredibly natural to our species. Every known human society has utilised spoken language for communication (Pinker & Bloom, 1990). Human infants need no formal instruction in the workings of their native tongue; they quickly become proficient simply by being exposed to people and conversations around them. This apparent ease with which we are able to comprehend spoken language nonetheless belies what is in actuality a complex and multi-faceted phenomenon, the underlying neural mechanisms of which we still do not fully understand. Scientists endeavouring to build speech recognition systems will attest to the complications of performing such a feat robustly – machines can now outperform us at various circumscribed tasks but achieving human-level speech understanding in a real-world environment remains elusive.

Speech comprehension is a necessarily intricate process in part due to the complexity of everyday acoustic scenes. In our daily lives, conversations seldom occur amidst a silent backdrop; the listener is quite often faced with making sense of their dialogue partner over a cacophony of competing sounds that impinge on their ears near-simultaneously. Colin Cherry (1953) coined the term 'cocktail party problem' to describe such a scenario, in reference to the difficult listening challenge that the eponymous event poses – multiple concurrent talkers, laughter and the clinking of glasses, the jazz band in the background. Most normal-hearing people are nonetheless still remarkably good at parsing such scenes, possessing the ability to understand their talker of choice as well as switch their attention at will to eavesdrop on any interesting chit-chat in their vicinity.

Background noise aside, construing meaning from an isolated speech stream is a challenge in itself. For one, there is acoustic variability that stems from inter-speaker differences – the same message articulated by any two talkers will have different acoustic features due to their individual voice characteristics and accents. Yet, human listeners are able to form a consistent percept of words despite these low-level spectrotemporal variations. Of course, simply understanding each word in isolation would still be insufficient to decode everyday speech. Meaning in human language is built up by the combination of words and phrases according to certain recursive rules – arguably the trait that most clearly distinguishes us from all the planet's other extant species. It is this unique structure of human language that affords us alone the ability to express an infinite number of ideas (Hauser, Chomsky, & Fitch, 2002). The parsing and integration of linguistic units is therefore also necessary to uncover the underlying message. Moreover, linguistic segments are not simply uttered at a fixed rate and tone. We thread them together with rhythms and melodies – collectively termed prosody – that inform how the speech is to be parsed (Steinhauer, et al., 1999) and communicate other layers of meaning. Given the strata of information contained in the speech signal, and the characteristic presence of background noise, our ability to comprehend spoken language in real-time is nothing short of remarkable.

The brain basis of speech comprehension – where and how cortex deals with the aforementioned issues – has been an area of inquiry for many decades. Yet, the answers to many fundamental questions remain elusive. Speech processing models posit that cortex is organised to process speech via two streams (Hickok, 2000; Hickok & Poeppel, 2004, 2007), and that the brain extracts meaning from the sound waveform via a hierarchy of increasingly abstract intermediary representations, with feedforward and feedback connections between the levels (Gaskell & Marslen-Wilson, 1997; Marslen-Wilson, 1987; McClelland & Elman, 1986). But many aspects of speech processing – e.g. the number and precise nature of stages required, and how selective attention influences the neural computations involved in processing sounds (and in particular, speech) – are still not well-specified. These are extensive questions that are practically beyond the reach of any one thesis; nonetheless, they form the broader context of this work.

1.1 Theme and motivations

The central theme of this thesis is how a particular approach – the use of electroencephalography (EEG) in conjunction with encoding/decoding methods – can be applied to tackle several research questions within the scope of the aforementioned topics. Many significant discoveries in cognitive neuroscience have only been made in the last few decades, not least because early researchers did not have any techniques at their disposal for glimpsing into the black box that is the living human brain. But the advent of modern neuroimaging techniques, allowing for the recording of brain signals *in vivo*, has opened up new doors. Still, each modality has its pros and cons, and selecting one is contingent upon the specific purpose for its use – in this case, to investigate speech processing dynamics in healthy humans. Electrocorticography (ECoG), grids placed on the surface of the brain, measures neural activity with high temporal and spatial resolution. However, it is invasive in that an incision has to be made through the skull for grid installation, which precludes its usage in healthy humans. Functional magnetic resonance imaging (fMRI), a non-invasive and spatially-precise measure of haemodynamic response, is a useful technique for localising speech-related areas in healthy and clinical subjects. But blood flow is sluggish relative to the rapid dynamics of speech, making fMRI less than ideal for isolating the individual stages involved in parsing speech. In this respect, electro-/magnetoencephalography (EEG/MEG), two modalities that measure the brain's electric/magnetic fields from the scalp, are excellent. They are able to record with millisecond precision, potentially allowing for the disentanglement of various speech processing stages in time.

In this thesis, EEG is used – its selection further motivated by its two additional advantages: it is portable and relatively cheap to acquire. These properties make it ideal as a clinical diagnostic tool for widespread use, and for interfacing the brain with machines. Indeed, while the scientific pursuit of understanding the brain basis of speech comprehension is fascinating in itself, the ability to pinpoint specific speech and language processing stages occurring in the brain using EEG could also lead to significant real-life impact for those with speech and language disorders and the hearing impaired. For one, it could facilitate earlier and more specific diagnoses, which would in turn aid the prescription of treatment or therapy. There is also increasing interest in incorporating EEG into next-generation, 'smart' assistive devices such as cognitively-controlled hearing aids (Lunner, Rudner, & Rönnberg, 2009). Hearing-impaired listeners struggle more so than usual in cocktail-party-like environments, when there are multiple co-occurring conversations. Current state-of-the-art hearing aids are able to isolate and suppress non-speech sounds based on acoustic statistics, but the capacity to decode the user's attended talker from their neural activity for amplification would be of even greater benefit. Recent studies on healthy subjects have found evidence for enhanced representation of the attended speaker in their neural activity (Ding & Simon, 2012; O'Sullivan et al., 2015). This discovery presents a potential solution for implementing 'smart' hearing aids: using attention-modulated EEG signatures of speech processing as control signals for selecting a speech stream. However, more research is necessary for it to be realisable in real-time and in a real-life setting.

A general limitation of EEG is its poor task-related signal-to-noise ratio due to contamination from muscular artefacts, unrelated brain activity, and external sources. To deal with this, early EEG studies had to present subjects with multiple repetitions – hundreds, typically – of simple, discrete stimuli. Averaging of their neural responses to a particular stimulus across all repetitions would then yield a visible temporally-detailed response to the stimulus at hand. But while this technique has been useful for identifying consistent patterns of responses across certain tasks, its findings do not necessarily translate to everyday speech processing. Humans typically deal with far more complex stimuli than repeated isolated syllables or words in the real world. Recently, the use of neuroimaging modalities in combination with encoding/decoding model-based methods (reviewed in Holdgraf et al., 2017) has been found to be an effective way of investigating how humans process more naturalistic stimuli. This entails presenting subjects with continuous stimuli that vary along multiple dimensions, as would occur in the real-world, and modelling their simultaneously-recorded neural signals using those features (or vice versa – using patterns in brain activity to predict/reconstruct features). This technique has been successful in estimating a temporally-detailed electrophysiological response to the amplitude envelope of speech (Lalor & Foxe, 2010). Moreover, in a simplified cocktail party scenario where subjects attend to one of two concurrent stationary talkers, it has been shown that both talkers' amplitude envelopes can be separately reconstructed from EEG data. Crucially, reconstruction of the attended talker's envelope is more accurate, providing a means for detecting the attended talker (O'Sullivan et al., 2015). This approach – envelope reconstruction – is currently the state-of-the-art for decoding selective auditory attention.

While the tracking of the envelope is a robust effect that can be leveraged for decoding attention and studying speech processing (e.g. by means of observing how it is affected by manipulations of the stimulus or cognitive states), there is perhaps an overemphasis on this phenomenon in the field. The envelope is a low-level measure of acoustic energy that covaries with many other acoustic features of speech, but it does not embody all information in the speech signal. There is therefore an opportunity for investigating how incorporating measures beyond the envelope can improve attentional decoding with a view towards real-time and real-world application. Furthermore, the estimated temporal response of this single aggregate feature is insufficient for disambiguating and characterising the precise computations involved in speech processing. In light of this, several studies in the last few years have sought to relate EEG activity to other features thought to be necessary as intermediate representations for the brain to decode meaning from the speech signal – inspired by the different levels of abstraction at which speech can be described in the parallel field of linguistics. These studies have found evidence that neural activity as indexed by EEG is best described as a combination of the acoustic features like the envelope and its onsets (Daube, Ince, & Gross, 2019) and higher-level speech representations like phonemes (Di Liberto et al., 2015), phonotactics (Di Liberto et al., 2018) and semantic markers (Broderick et al., 2018). Nonetheless, these features still do not completely characterise the path from sound to meaning, and there remain representations essential for understanding speech which have not been explored using EEG.

1.2 Aims

The overarching aim of this thesis is to increase our overall understanding of the neural mechanisms underpinning speech comprehension through the employment of encoding/decoding methods on EEG data to isolate neural signatures of speech processing beyond the response to the acoustic envelope, with a view towards real-world application. Within this context, we tackle three specific aims:

a) to investigate a recently-established auditory attention decoding framework in a cocktail party scenario with non-stationary talkers – as would not be uncommon in real life – and improve upon its efficacy by incorporating measures beyond the envelope;
b) to isolate and characterise EEG indices of the cortical processing of prosody; and
c) to explore the effects of selective auditory attention on pre-lexical and prosodic speech processing.

1.3 Outline

In Chapter 2, our current understanding of the neural basis of speech processing is reviewed – including the neuroanatomy of the auditory system, the functional organisation of cortex for speech processing, and the role of top-down attention in parsing auditory scenes. EEG as a neuroimaging technique and the encoding/decoding approach employed in this thesis is also discussed, with a particular focus on how it has been applied to EEG data so far within the context of speech processing and the decoding of selective auditory attention.

In Chapter 3, the robustness of an established state-of-the-art framework for decoding auditory attention (O'Sullivan et al, 2015) is investigated within the context of a paradigm with non-stationary speakers. The incorporation of other EEG indices of attention that are practically computable in real-time into the algorithm – namely, the model weights themselves, and alpha power – is also considered in the effort to improve decoding performance.

In Chapter 4, EEG responses to measures of prosodic pitch – relative pitch and resolvability – are explored. The encoding of relative pitch has been observed in ECoG data (Tang et al., 2017), and cortical regions that respond primarily to resolved harmonics have been found using fMRI (Norman-Haignere et al., 2013). Nonetheless, it remains to be seen whether EEG is sensitive enough to capture neural correlates of the cortical processing of pitch.

In Chapter 5, we investigate how attention modulates EEG signatures posited to represent processing at the pre-lexical level, as well as indices of prosody. The former inquiry is motivated by the longstanding debates as to whether attention modulation operates at an early or late stage in the speech processing hierarchy, and whether an acoustic-phonetic transformation stage occurs as an intermediate step in mapping the acoustic speech signal to the mental store of words (lexicon); these outstanding issues are further elaborated upon in Sections 2.3.4 and 2.4 of the literature review.

In Chapter 6, a general discussion of the work in the thesis and its place within the broader field is provided.

Chapter 2 Literature Review

As established in the previous chapter, there is still much we do not know about how humans comprehend everyday speech. Nonetheless, the last few decades in particular have seen a lot of progress towards increasing our understanding of this phenomenon. This chapter details pertinent background information and summarises the literature in several subareas relevant to the thesis. To start with, a cursory overview of human anatomy relevant to speech processing – the cerebral cortex (from which signals recorded using EEG primarily emanates) and the auditory system – will be provided. Next, the path from sound to meaning – how cortex is functionally organised to extract meaning from speech and precisely what computations might be necessary based on linguistic observations – will be discussed. Auditory scene analysis, which concerns how we hear in noisy environments, will then be reviewed.

This chapter also reviews EEG as a neuroimaging technique. An explanation of encoding/decoding methods, with emphasis on the temporal response function (TRF) approach that will be employed throughout this thesis, will be provided.

2.1 The cerebral cortex

The cerebral cortex is the superficial grey matter portion of the human brain that concerns itself with higher-order processes. Its surface is highly convoluted – it consists of grooves called sulci, and elevated ridges between them called gyri. The cerebral cortex is made up of two hemispheres – left and right – that are separated by a deep groove, the medial longitudinal fissure. Each hemisphere is further subdivided by fissures into four distinct sections, or lobes: frontal, parietal, occipital, and temporal (Figure 2-1). Within each lobe of the cortex, there are numerous cortical areas, each determined by its functionality. Auditory cortex – the area for processing auditory information, which is of particular relevance to this thesis – is part of the temporal lobe.

Sensory areas such as auditory cortex can be further subdivided into regions. Primary areas are those that receive direct sensory input from the ascending sensory pathways. Interconnected with primary sensory areas are secondary sensory regions – they analyse the signals passed on by the primary areas. There are also motor areas that concern themselves with the voluntary contraction of muscles. Lastly, there are areas of the cortex not associated with sensory or motor function – the association areas. These areas intervene between sensory and motor areas and support higher cognitive processing such as learning, decision making, attention and – of particular interest in this thesis – language abilities. The subareas involved in speech and language will be described in greater detail in the subsection reviewing models of speech processing.

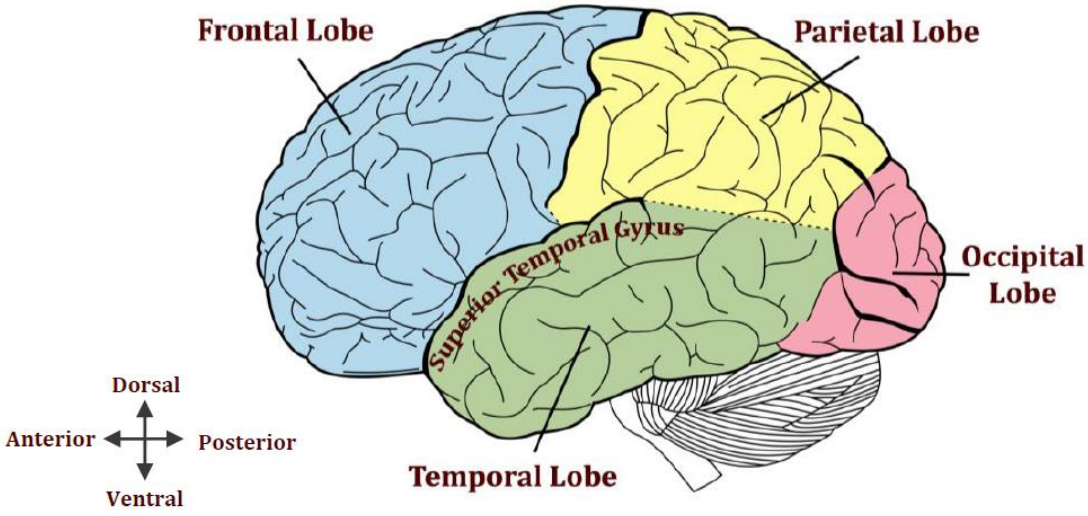


Figure 2-1 | The cerebral cortex.
Lateral view of the human cerebrum showing the four lobes of the cerebral cortex (left hemisphere, but right hemisphere also consists of the same four lobes). The auditory cortex is located on the superior temporal gyrus, which is indicated in the figure (Adapted from Gray & Lewis, 1918).

2.2 The auditory system

The auditory system is the sensory system for hearing. It can broadly be divided into two subsystems: peripheral and central. The anatomy of the auditory periphery consists of the outer ear, middle ear and inner ear (Figure 2-2). Collectively, these parts function to efficiently transduce sound – mechanical vibrations of the surrounding medium, typically air – into neural action potential signals. The acoustic information, now in electrical form, travels through the ascending auditory pathway and is relayed to primary auditory cortex for further processing. These structures make up the central auditory system.

2.2.1 Auditory periphery

The outer ear, made up of the pinna and auditory canal, collects sound from a wide area. These vibrations impinge upon the tympanic membrane (i.e. eardrum) of the middle ear, causing it to vibrate. The movements of the membrane are then transferred through a series of bones called the ossicles into movements of a second membrane, the oval window. This window seals the top chamber of the cochlea.

The top and bottom chambers are filled with incompressible fluid called perilymph. The middle channel holds a complex structure called the organ of Corti, which sits on a membrane known as the basilar membrane. It is this structure that serves as the interface between the peripheral and central auditory subsystems. Hanging over the organ of Corti is a second membrane called the tectorial membrane. The organ of Corti contains rows of hair cells with tiny cilia that protrude across their surface. When the oval window vibrates, it sends a travelling wave within the cochlea. This in turn causes the basilar membrane, and therefore the organ of Corti and its hair cells, to move up and down, towards and away from the tectorial membrane. When the basilar membrane deflects upwards, the cilia on the hair cells become bent from being pushed against the tectorial membrane. This movement causes the generation of neural signals in the hair cells. These signals are sent to the brain through the cochlear branch of the auditory-vestibular nerve. The more a particular section of the basilar membrane vibrates, the higher the firing rate of the auditory nerve fibres that come from this patch.

The cochlea transduces sound according to frequency. The distance that the wave travels up the basilar membrane is dependent on the frequency content of the incoming sound. High frequencies cause the stiff base of the membrane to vibrate more, dissipating the energy before it can be transmitted further. Low frequency sounds on the other hand travel further up to the apex before most of the energy is dissipated. The consequence of this is that there is a corresponding representation of frequency in the auditory nerve, and this spatial arrangement of frequency, called tonotopy, is maintained throughout the classical auditory pathway to the level of the primary auditory region.

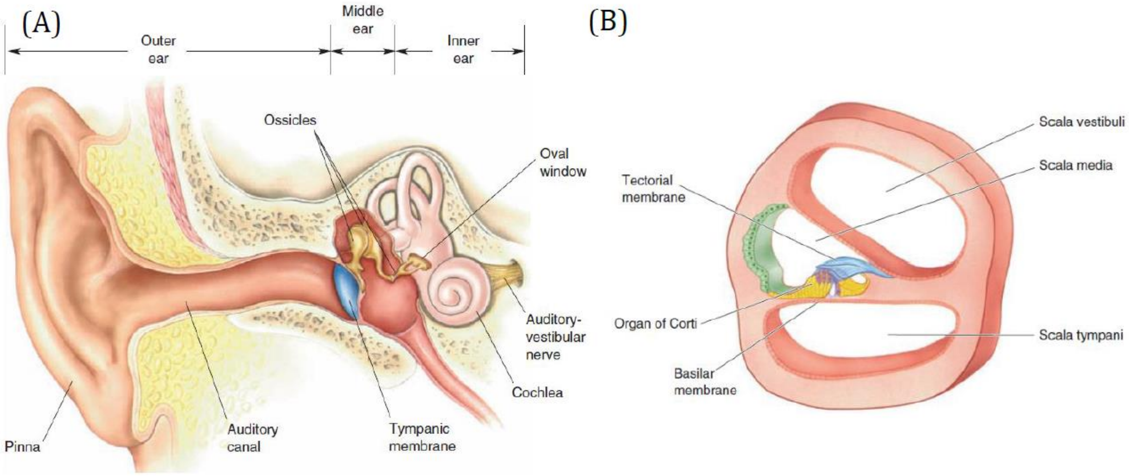


Figure 2-2 | The auditory periphery

(A) The auditory periphery consisting of the outer ear, middle ear and inner ear (adapted from Bear et al., 2007) (B) Cross section of the cochlea showing the three chambers (adapted from Bear et al., 2007)

2.2.2 Central auditory system

The central auditory system is made up of ascending and descending pathways. The pathways of the auditory system are arguably more complex than those of other somatosensory systems and remain incompletely characterised. There are two parallel ascending pathway systems – classical and non-classical – that process information differently and have different central targets. In the non-classical pathway system, neurons are not as clearly tuned, and input is also received from other sensory systems. This system bypasses primary cortical areas and projects to secondary areas (Møller, 2013). In the more well-known classical pathway system, auditory nerve axons innervate the dorsal cochlear nucleus (CN) and ventral CN in the medulla oblongata of the brain stem. Fibres (called the lateral lemniscus; LL) then cross over to the opposite side and connect to the central nucleus of the contralateral inferior colliculus (ICC) of the midbrain, a nucleus that acts as an integration station and switchboard. From there, the ascending auditory sensations synapse in the medial geniculate nucleus/body (MGN/MGB) of the thalamus before being relayed to cortex.

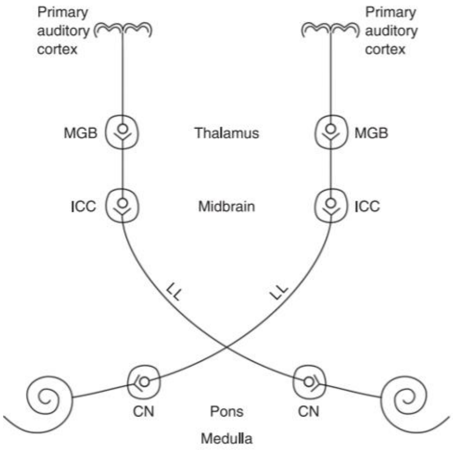


Figure 2-3 | Schematic diagram of main nuclei and fibre tracts of the classical ascending auditory system pathway (Møller, 2013)

The thalamo-cortical projection of the system consists of a highly myelinated group of axonal projections termed the acoustic radiation (AR). The AR takes a medio-lateral course from the MGN to primary auditory cortex. The available anatomical information on this white matter bundle in humans remains limited. It is a challenge to reconstruct this bundle *in vivo* using diffusion tractography due to its small size and location. Nonetheless, precise characterisation would reveal whether there is any volume lateralisation at this level, where left lateralisation may suggest some language-specific encoding even at this level (Maffei et al., 2019).

We know relatively little about the functional organisation of human auditory cortex when compared to what we know about visual cortex. A lot of auditory cortical models rely on projections from studies in non-human animals. Primary auditory cortex, or A1, is located in Heschl's gyrus and the superior temporal gyrus (STG) in the temporal lobe (depicted in Figure 2-1 and Figure 2-4, but STG's location is subject to hemispheric and individual variability). A1 receives point-to-point input from the ventral MGN and therefore maintains a tonotopic arrangement of frequencies (Da Costa et al., 2011). Beyond tonotopy, there is also some evidence for tuning in A1 for low-level features such as temporal and spectral modulations (Herdener et al., 2013; Schönwiesner & Zatorre, 2009).

Surrounding A1 and interconnected with it are secondary or belt areas. These areas receive projections from MGN (mainly dorsal) via the non-classical pathway and from A1. They have been shown to be selective for more complex stimuli as opposed to simple pure tones (e.g. noise bursts; Wessinger et al., 2001). This is consistent with the idea that auditory processing is hierarchical, where higher areas process increasingly complex/abstract information. In humans, the precise division and configuration of these areas and their functional significance remain incompletely investigated. Studies using controlled stimulus generated specifically to test for responses to individual acoustic dimensions of sound have found selectivity in non-primary auditory cortex for the processing of features like pitch, the perceptual correlate of sound periodicity (Penagos, Melcher, & Oxenham, 2004).

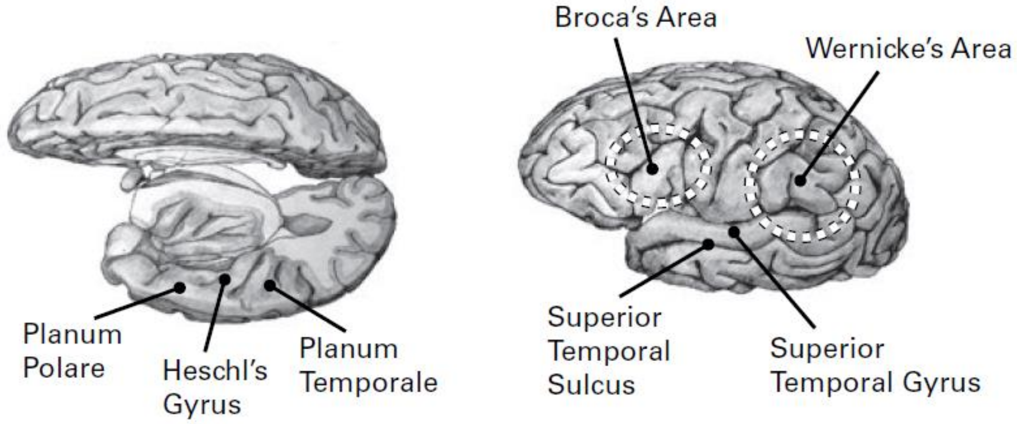


Figure 2-4 | Cortical speech processing areas

(L) Top-view of a transverse slice of left hemispheric cortex showing where the primary auditory (superior temporal/Heschl's gyrus) and belt areas (planum temporale and polare) are situated within auditory cortex. (R) Lateral view of left hemispheric cortex showing other areas commonly associated with higher-order auditory processing and speech (Adapted from Schnupp, Nelken, & King, 2011)

Other areas in human cortex posited to be important for the processing of higher-order auditory stimuli and speech are superior temporal sulcus (STS) and Broca's and Wernicke's areas, shown in Figure 2-4. Broca's and Wernicke's areas are named after nineteenth century neurologists, who discovered through lesion studies that damage to these areas were associated with disruption to speech production and comprehension, respectively. We will examine STS in greater detail in the next section within the context of the dual-stream model for speech processing.

Although we have largely focused on feedforward connections, there are also descending projections relaying information back down from cortical regions all the way down to cochlea. Indeed, auditory processing involves an interplay of bottom-up and top-down signals. For example, ‘olivocochlear bundle’ axons originate in the superior olivary complex (SOC) in the brainstem, and travel with the auditory vestibular nerve to synapse either directly with the outer hair cells or on the endings of the auditory nerve fibres that innervate the inner hair cells in the cochlear (Schnupp et al., 2011). The descending pathways in the auditory system are thought to be as abundant was ascending pathways, although much less is known about them. Nonetheless, their presence indicates that auditory processing likely incorporates feedback loops on many levels. This makes it possible for the auditory system to be re-tuned on the fly to meet the demands of a particular task or environment.

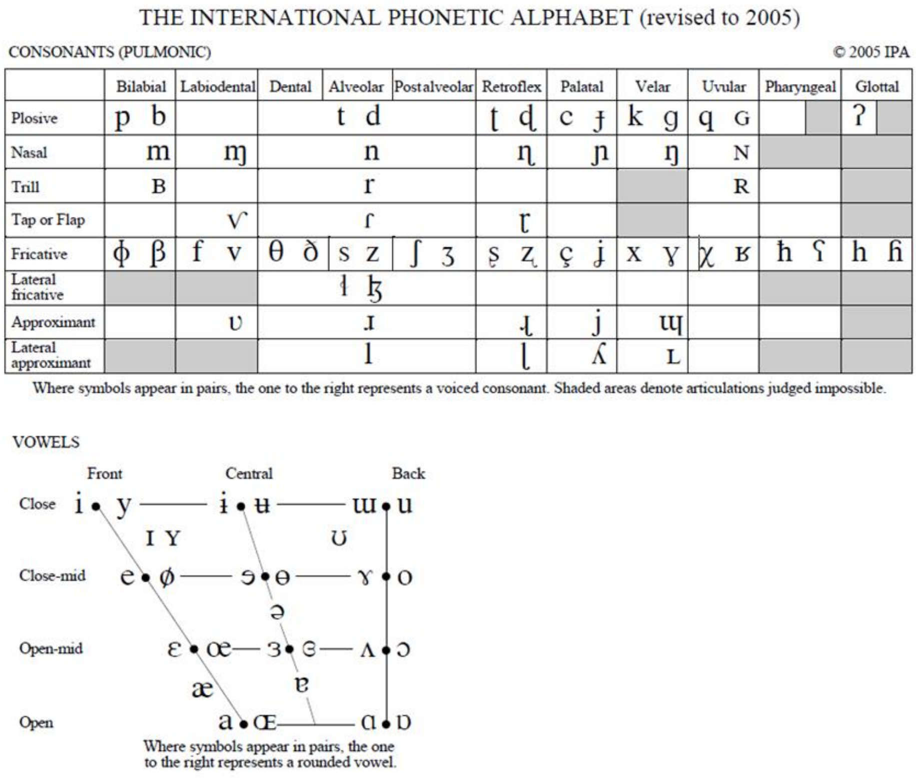
2.3 From sound to meaning

EEG is used in this thesis and its poor spatial resolution precludes detailed anatomical localisation; nonetheless, understanding the functional neuroanatomy of speech processing is important for placing this work within the larger field. In this section, we describe the dual-stream model (Hickok & Poeppel, 2000, 2004, 2007), an influential contemporary theory that integrates empirical findings to outline how cortex might be organised to support speech perception and linguistic function. We refer to speech perception here as the processes involved in transforming the auditory signal into a representation that accesses the mental store of words, and linguistic function as the higher-order language processes beyond that. To better understand the representations that cortex might compute in going from sound to meaning, we first briefly review observations from linguistics on the structure of speech and how meaning is built up.

2.3.1 Linguistic structure of speech

In order to speak and comprehend speech, every person who knows a language has internalised its system of rules that determines its sound-meaning connections (Aitchison, 2007; Chomsky & Halle, 1968). Research in linguistics hypothesises about this internalised system based on examining the speech that humans produce as a natural object in and of itself.

Starting with the repertoire of sounds used in any given language, linguists have surmised that they only differ from each other in a small number of ways in terms of how they are articulated. The set of these differences have been termed distinctive (or phonetic) features (Chomsky & Halle, 1968; Jakobson & Halle, 1956). They include properties of consonants such as place of articulation (the location within the oral cavity where the constriction and obstruction of air occurs), manner of articulation (the configuration and interaction of speech organs such as tongue, lips, and palate), and laryngeal features that specify the glottal state of sounds, as well as properties of vowels (e.g., back, high, and tense vowels). These features form the smallest categorical units of speech that also have an acoustic interpretation. They provide a connection between action and perception in speech and are the basic inventory characterising the sounds of all languages – that is, they are not language-specific (Hickok & Poeppel, 2007). Bundles of coordinated phonetic features that are concurrent in time form segments or phonemes. Each language has its own set of phonemes which are in turned organised into syllables. Constraints apply with regards to how phonemes can be arranged within a syllable or at syllabic boundaries. In English, a syllable is a unit of sound made up of a vowel sound, with or without surrounding consonants. These three representations – features, phonemes, and syllables – provide the infrastructure for pre-lexical phonological analysis.



THE INTERNATIONAL PHONETIC ALPHABET (revised to 2005)

CONSONANTS (PULMONIC) © 2005 IPA

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k ɡ	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap		ⱱ		ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

VOWELS

Where symbols appear in pairs, the one to the right represents a rounded vowel.

Figure 2-5 | International Phonetic Alphabet (IPA)

A subset of phonemes and their relation to underlying phonetic features according to the International Phonetic Alphabet (IPA) convention (Adapted from Gussenhoven & Jacobs, *2013). The IPA was devised to be a standardized representation of the sounds of spoken language.*

Beyond pre-lexical representations are linguistic units that contain meaning/semantics, with morphemes traditionally regarded as the smallest of these structures. One or more morphemes are combined to make up words – the difference between the unit types are that words can always occur freely, but some morphemes cannot. The word ‘jumped’, for example, consists of the free morpheme ‘jump’ and the bound morpheme ‘ed’ (Spivey, Joanisse, & McRae, 2012). Morphemes and words often have more than one lexical meaning – in which case their interpretation becomes dependent on the surrounding context. Words are assembled to form phrases in accordance with certain rules called syntax. Each word has a ‘part-of-speech’ category (e.g. noun, verb, adjective) which constrains its role and determines where it can appear in building up a phrase. Phrases themselves fall into several categories (e.g. noun phrase, verb phrase) which can be combined and manipulated in chunks to form sentences that convey larger ideas.

While linguists generally agree upon the aforementioned hierarchical levels at which language can be described, there are some differences between linguistic theories on the categories within each level. Additionally, it is often also not straightforward to analytically parse everyday speech into these organisational structures. Natural speech is riddled with ambiguity – in the way words are pronounced by different speakers (there are no known invariant acoustic cues across instances of a phoneme), in the meaning of words, in the way words are assembled to form a sentence and so on – such that context has to be taken into account in order to parse speech into its various constituents. Computational methods have been developed in recent years to automatically parse speech at various levels – these tools are still unable to rival the ability of humans in recognising and construing meaning from speech, but they enable us to investigate if the proposed organisational levels and categories might correspond to actual representations that exist in the brain.

Apart from phonetic, syntactic, and semantic content, speech allows for another layer of meaning to be transmitted via prosody; that is, the variation of acoustic cues that tie linguistic segments together. Prosody is characterized by phenomena such as intonation, stress, and rhythm. The same sequence of words can convey very different meanings and emotions by virtue of prosodic differences (Bolinger, 1986); e.g., the intonation contour of a sentence a can determine if someone is asking a question or making a statement. Additionally, prosody also aids in disambiguating syntactic constituent structure. Prosodic phenomena are manifested through the modulation of acoustic percepts like pitch, loudness, and duration. For example, stress, which is the prosodic event of extra emphasis being placed on a segment, can be produced in English by using higher pitch, and/or increasing the sound's duration or loudness (Chrabaszcz, et al., 2014), and intonation is typically conveyed by the variation of pitch across segments. A prosodic percept can arise from the modulation of one or multiple acoustic measures and it remains unclear to what extent individual cues or cue combinations contribute to a particular prosodic feature. Thus far, systems for identifying and characterising prosodic elements and how they map to meaning are less established (Cole & Shattuck-Hufnagel, 2016).

2.3.2 The dual-stream model for speech processing

The dual-stream model outlines processing starting at the level of primary auditory cortex, where spectro-temporal analysis of auditory signals received from the thalamus is conducted. According to the model, the early cortical stages of speech perception are organised both hierarchically and bilaterally. Some evidence has been found to support this. For example, in an fMRI study, Binder et al. (2000) presented participants with five types of auditory stimuli: unstructured noise, frequency modulated tones, words, pronounceable pseudowords, and temporally-reversed words. They found the following activation patterns bilaterally: the dorsal plane of the STG responded more to tones than noise, and regions further downstream in mid-STG showed selectivity for speech over tones.

After the initial cortical stages of speech perception, the model posits that processing is divided into two computational pathways – a largely bilateral ventral one for auditory-conceptual processing, and a left-dominant dorsal one for auditory-motor processing. The neuroanatomical regions that make up these pathways are depicted in Figure 2-6. The dual-stream aspect of the model was put forth to resolve the paradoxical results of previous studies in which it was observed that passive listening tasks implicated superior temporal regions bilaterally (Binder et al., 1994; Mazoyer et al., 1993) but studies that used tasks similar to syllable discrimination or identification found prominent activations in the left STG and left inferior frontal lobe (Zatorre, et al., 1992).

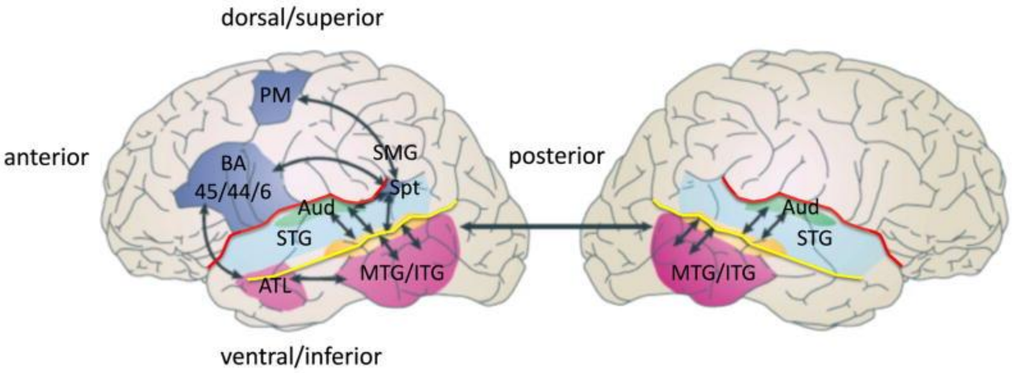


Figure 2-6 | The dual-stream model

Functional neuroanatomy of the dual-stream model for speech processing (from Hickok, 2013). A1 (Aud) and STG are involved in the initial stages of speech perception. Processing then diverges into two pathways: the dorsal pathway (dark blue areas) conducts auditory-motor processing while the ventral pathway (pink areas) is for auditory-conceptual processing.

The ventral stream can be broadly understood as the ‘what’ pathway for auditory processing. It is posited to consist of a ‘lexical interface’ for mapping the sound/phonological structures of words onto the corresponding semantic representations, and a ‘combinatorial network’ for forming the integrated meanings of phrases and sentences (Kemmerer, 2014). The posterior, middle, and ventral sections of the temporal lobe (middle and inferior temporal gyri, MTG/ITG in Figure 2-5) appear to be important components of the lexical interface that serve as the bridge between sound and meaning. Lesion studies have shown that damage to these regions results in semantic-level deficits in both comprehension and production (Chertkow et al., 1997; Hickok and Poeppel 2004, 2007). fMRI studies have also implicated these regions for lexical-semantic processing – Binder et al. (1997) found that when subjects made semantic decisions about auditory words, portions of the STS and MTG/ITG were strongly activated (in addition to frontal and parietal regions). STS activation was interpreted to be due to the phonological aspects of word processing, and MTG/ITG to reflect post-phonemic mechanisms involved in processing or accessing lexical-semantic information (Hickok, 2013). The lexical interface region is posited to project forward to lateral portions of the anterior temporal lobe (ATL), especially in the left hemisphere. There is preliminary support from fMRI studies indicating that the ATL region houses a ‘combinatorial network’ for processing compositional meaning – e.g. Narain et al. (2003) found stronger responses in this area to intelligible than unintelligible multi-word utterances and Vandenberghe et al. (2002) found portions of left ATL to be more active when subjects read normal sentences as opposed to scrambled ones. However, more work still needs to be done to flesh out the processes involved in parsing phrases and sentences.

The dorsal stream can be broadly thought of as the 'how' pathway of the auditory system. It is posited to consist of a 'sensorimotor interface' that has reciprocal connections with the phonological network, and a 'articulatory network' that has reciprocal connections with the sensorimotor interface (Kemmerer, 2014). Auditory representations that arrive after phonological processing are sent to area Spt, located in the posterior-most part of the left sylvian fissure. This area conducts a sensorimotor transformation of the input and transmits the signals forward to the articulatory network in the left posterior frontal lobe. This circuit is thought to contribute to several aspects of the human capacity for language, including the acquisition of speech-motor patterns and the perceptual processing of speech, particularly in situations when the listener must pay close attention to the phonological structure of utterances – a potential reason why it is activated in tasks like syllable discrimination or identification but not during passive listening.

The dual-stream model outlines a general framework based on empirical evidence for the cortical organisation of speech and provides context for speech and language processing research. Nonetheless, it is broad in scope and focuses on the anatomical aspect rather than the precise computations involved and time courses of the various processes. In Section 2.6.4, we provide a review on recent research that use temporally-precise neuroimaging methods such as EEG and MEG to shed more light on these areas.

2.3.3 Prosodic processing

The processing of prosodic cues is not addressed in Hickok and Poeppel's dual-stream model. Nonetheless, there have been various studies that have sought to anatomically localise prosodic processing in the brain, with hemispheric lateralisation – whether or not there is right asymmetry – being a point of contention (Baum & Pell, 1999; Gandour et al., 2004; Kreitewolf et al., 2014; Meyer et al., 2002; Plante et al., 2002; Witteman et al., 2011). Recently, there has also been an attempt to devise a network model for prosodic processing akin to Hickok and Poeppel's model. Sammler et al. (2015) conducted a study in which subjects listened to words that gradually varied across phonemic ('bear' to 'pear') and prosodic (question to statement) dimensions and recorded their neural responses using fMRI. Subjects were asked to perform one of two tasks after listening to each word – a prosodic one where they categorised the word as a question or a statement, and a phonemic one where they categorised the word as 'bear' or 'pear'. By contrasting responses during the two tasks, they identified stronger activations in the right posterior (pSTS) and anterior (aSTS) superior temporal sulcus, the right inferior frontal gyrus (IFG), and the right premotor cortex (PMC) during the prosodic task. They then used these activation clusters as seed and target regions in multi-fiber probabilistic tractography to estimate the most likely white-matter pathways that connect these prosody-relevant nodes. This approach revealed a ventral tract connecting pSTS and aSTS, and dorsal tracts connecting pSTS to IFG and PMC.

In light of these results, as well as the findings of earlier functional connectivity studies showing parallel pathways for prosody (Ethofer et al., 2006), they proposed that prosodic processing also proceeds along two streams, albeit with rightwards asymmetry. The dorsal pathway in this case consists of regions in right PMC and IFG – they are posited to convert prosodic contours into a motor format for articulation. The ventral pathway, comprising regions in STS, is posited to perform gradual segregation and abstraction of the prosodic signal from a speaker-dependent representation of speech sounds to a speaker-invariant representation. That is, acoustic cues relevant for conveying prosodic information are thought to be computed and integrated within this pathway.

Nonetheless, a comprehensive model of how the brain processes speech also necessitates understanding precisely how this prosodic information is encoded in the brain. This is complicated by the fact that there is no established standard for the representation of prosodic elements that generalises across languages (Hirst, 2005). But an undeniably important auditory percept for prosody, the modulation of which underlies various prosodic phenomena, is pitch, which can be estimated from the acoustic speech stream (Boersma, 1993). With regards to this, it has been established that there are auditory cortical regions that encode sound pitch (Griffiths et al., 2010; Hall & Plack, 2009; Briley et al., 2013), and that these pitch-sensitive regions respond primarily to resolved harmonics (Norman-Haignere et al., 2013). In the context of speech, evidence has recently been found for the cortical encoding of speaker-normalized relative pitch. Tang et al. (2017) measured high-gamma electrocorticography (ECoG) responses to set of spoken sentences that independently varied intonation contour, phonetic content, and speaker identity, and found high-gamma activity on single electrodes over superior temporal gyrus selectively represented intonation contours in terms of relative pitch (Figure 2-7). This demonstrates how intonation undergoes specialized extraction from the speech signal, distinct from other aspects of speech. An outstanding future question concerns characterising how prosodic and segmental information are integrated to support a unified percept for language comprehension.

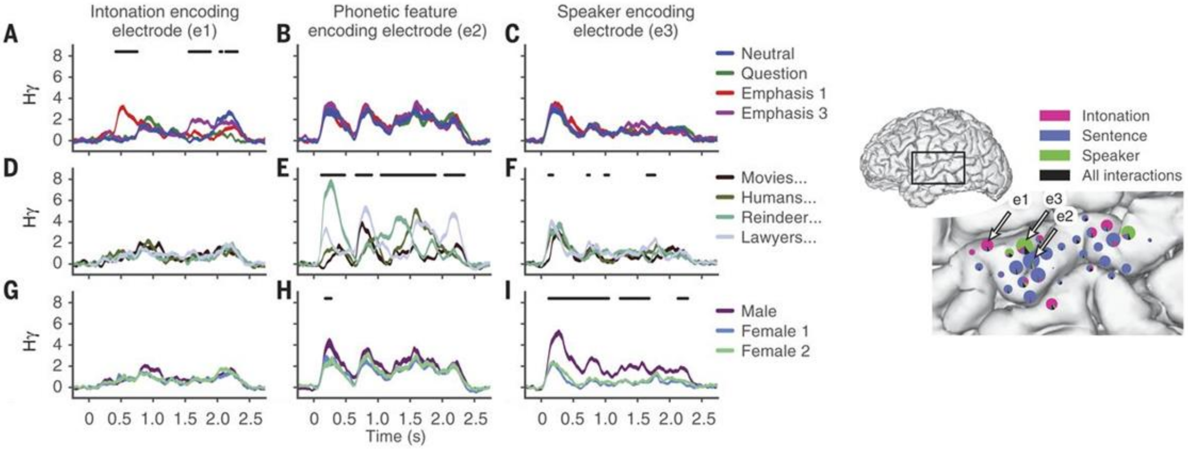


Figure 2-7 | ECoG activity displays selectivity for intonation

Neural responses at three example ECoG electrodes that are selective for three different speech features – intonation, sentence (i.e. linguistic units), and speaker (adapted from Tang et al., 2017). The locations of these electrodes are indicated in the subplot on the right. The area of the pie occupied by a particular feature at each location is proportional to the total variance explained by the feature. Subplots A to C show the responses at these electrodes to a sentence spoken by a single talker but with different intonation contours. It can be seen that neural activity on electrode one (A) differentiates intonation contours, whereas activity on electrodes two (B) and three (C) does not. D to F shows neural responses at those same electrodes as sentence content is varied but the other features are kept constant, whilst G to I shows neural responses as speakers are varied with other features kept constant.

2.3.4 Pre-lexical processing: the role of the phoneme

While linguistic theories and many speech processing models – including the widely-accepted dual-stream model discussed above – posit a mapping to phonetic features and/or phonemes as an intermediate stage between low-level acoustics and words (McClelland & Elman, 1986; Liberman et al., 1967; Hickok & Poeppel, 2007), this viewpoint is not unanimous. Some researchers question the necessity and existence of a mental representation at this level for comprehending speech meaning (Lotto & Holt, 2000). They point out that structure and organization in behaviour or output need not imply that this structure and organization is present in mental representation. There have been theories advocating for alternatives, including a different intermediate representational unit (e.g. syllables) (Massaro, 1974), or direct matching to the lexicon based on the comparison of acoustic representations (Goldinger, 1998; Klatt, 1989). A main criticism of a phoneme-level representation is that there is weak physical basis for phonemes – thus far, there has been a failure to find acoustic invariances across instances of a particular phoneme, and there is no straightforward way to map from the continuous acoustic speech signal to a phoneme representation without taking contextual constraints into account.

There has been evidence from behavioural and EEG studies that humans perceive phoneme categories. Liberman et al. (1957) found that discrimination within phoneme categorical boundaries is poorer than across them. Eimas et al. (1971) looked at infants' ability to discriminate syllables which differed in a voicing (phonetic) feature and found that infants also perceived categorical phonetic features. Additionally, Näätänen et al. (1997) found that the mismatch negativity (MMN), an event-related potential (ERP; see Section 2.5) component which reflects discriminable change in some repetitive aspect of the ongoing auditory stimulation, is observed when the deviant stimulus is a phoneme prototype of a subject's native language relative to when it is a non-prototype. Nonetheless, it has been argued that these studies used simple isolated units, and therefore their results could be due to the type of task and not reflective of everyday listening.

Recently, studies have shown that a phonetic feature representation of speech can predict neural activity during listening to natural, continuous speech. Using high density cortical surface electrodes, Mesgarani et al. (2014) found that the superior temporal gyrus (STG) (Figure 2-4) selectively responds to phonetic features. Di Liberto et al. (2015) and Khalighinejad et al. (2017) have also found evidence that non-invasive EEG activity reflects the categorisation of speech into phonetic features. Nonetheless, Daube et al. (2019) argue that these results could be also be explained by an encoding model that is based entirely on acoustic features. Namely, they found that acoustic onsets made very similar predictions to the benchmark phonetic features, and that acoustic onsets explained parts of the neural response that phonetic features could not. Contrasting evidence aside, the idea of mapping to words via abstract sub-lexical representations like phonemes has prevailed due to its computational efficiency (Scharenborg, Norris, Bosch, & McQueen, 2005). It is less obvious how the detection of acoustic onsets and other acoustic representations would lead to lexical access – and in particular, how such a model would robustly generalise to new words and new speakers. However, given that current evidence is insufficient to definitively conclude the precise computations the pre-lexical level, more research within this subtopic needs to be conducted.

2.4 Auditory scene analysis

Auditory scene analysis is concerned with how humans hear in noisy environments – a situation that many of us deal with on a daily basis. In spite of the world being noisy, humans are somehow able to detect patterns in the auditory scene to parse it into isolated components that occur simultaneously but can be attended to separately.

2.4.1 The cocktail party effect and early versus late selection theories

The phenomenon of attending to a single talker amidst multiple competing sources was brought to the fore by Colin Cherry in 1953 when he took an interest in it and coined the term ‘cocktail party effect’. The ‘cocktail party effect’ is essentially a speech-specific occurrence of auditory scene analysis, typically discussed with a focus on the attentional aspect. Cherry conducted a dichotic listening experiment where distinct speech streams were played to the left and right ears of subjects and they performed various tasks. His primary motivations were the questions of how humans are able to recognise what one person is saying when others are speaking concurrently, and how one might go about designing a machine to carry out such an operation. In his seminal paper, he found that subjects recognised certain low-level aspects of the unattended speech stream (e.g. they were aware if the speech was time-reversed) but were unable to recall any higher-level linguistic content. Cherry’s findings inspired more studies along the same vein – with a focus on how attention modulates the processing of speech in multi-talker scenarios.

A prominent theory of selective auditory attention from this period is Broadbent’s filter model of attention (1958). In his early-selection model, Broadbent assumes some initial automated bottom-up (i.e. stimulus-driven) processing of sounds, after which there is a strict filter that rejects sounds based on low-level features such as location, pitch, loudness, and timbre. A subsequent study (Moray, 1959) found that higher level information such as one’s own name is perceived even if it is in the unattended stream.

This prompted others to suggest modifications/alternatives to Broadbent's theory. Deutsch & Deutsch (1963) put forth a late-selection theory suggesting that all competing speech streams are semantically analysed, but only the attended stream is written to long-term memory. Treisman (1964) proposed a less drastic version of Broadbent's model in which unattended stimuli are attenuated rather than blocked out in an all-or-nothing fashion and therefore capable of reaching awareness in exceptional situations. The issue of whether attention to speech operates at an early stage of processing based on the physical characteristics of the stimulus or at a later stage during semantic processing is still debated, and Broadbent's filter/bottleneck metaphor remains influential on modern-day thinking on how attention operates.

Following these early investigations, selective attention studies largely turned to the visual domain and the role of attention. At the time, it had just been shown that there are areas in (primate) visual cortex tuned to basic features of the visual scene such as lines orientation and colour (Hubel & Wiesel, 1968), so the question of how a unified percept of an 'object' could arise from these isolated low-level features was of much interest. Treisman and Gelade (1980) proposed a feature-based theory of visual attention, according to which when attention is directed to one or more elements of a perceptual visual object, all its other low-level visual features (e.g. colour and shape) are also registered pre-attentively and are bound together. Meanwhile, auditory research during the 1970s and 80s was mostly concerned with how physical cues of sound are coded in the auditory periphery, with little regard for perceptual phenomena or the effects of attention. It was not until seminal work by Bregman (1990) that auditory neuroscience began focusing on auditory scene analysis again, albeit with emphasis on a different aspect having been inspired by vision studies.

2.4.2 Auditory object formation

Apart from the attentional aspect of auditory scene analysis, the listener has to deal with another conceptually distinct challenge in parsing their environment: forming auditory objects. This entails separating the mixture of sounds into their constituent sources and grouping them. In practice, the two challenges are intertwined for the listener – they occur heterarchically and influence one another (Shinn-Cunningham, Best, & Lee, 2017). But while the aforementioned early studies (e.g. Cherry and Broadbent) focused on selective attention, Bregman (1990) and the many studies building upon his work were concerned with the perceptual organisation (or grouping principles) of sound mixtures; namely, its properties and influencing factors. Research bridging the gap between these two lines of inquiry will be reviewed in Section 2.4.3, but in this subsection, we review the literature on auditory object formation.

Consider that the signal that impinges upon the ear is a mixture of all sounds present in the environment. Humans are nonetheless able to group sounds into perceptual components that are typically referred to as auditory objects (a term inspired by vision studies). To illustrate the idea of an auditory object, Bregman gives the example of an infant being spoken to by their mother. When the infant attempts to imitate their mother's voice, they do not insert the squeak from the cradle or other environmental sounds into their imitation; they are somehow able to distinguish that it is not part of the same perceptual object. The exact theoretical definition of an auditory object remains debated (see review by Griffiths & Warren, 2004), but it can be broadly thought of as a distinct perceptual time-frequency image attributed to a sound source; e.g. speech from a talker, or the barking of a dog. Segregating and grouping sounds into objects is non-trivial because determining how many sources there are in a given mixture and what energy components can be attributed to each source are ill-posed mathematical problems. To make these estimations, the brain has to utilise its implicit knowledge of specific sounds or sound classes to some extent (McDermott, 2009).

Auditory object formation is thought to occur at two time-scales (Carlyon, 2004; Shinn-Cunningham et al., 2017): a 'local' scale (approximately on the timescale of a speech syllable) to bind together sound energy components that are concurrent (sometimes called tokens), and a longer scale in which interleaved sound tokens from a mixture of sources are sorted into auditory objects that extend through time called streams (Bregman, 1990). How this grouping is accomplished can be described in terms of Gestalt principles, rules that have been formulated to describe how perceptual scenes are organised. These principles were first described in relation to vision but are also applicable to audition. At the shorter time scale, the rule of proximity in the spectro-temporal domain (Bregman 1990) is thought to be a basis for grouping. That is, sound components that are close together and continuous in time and/or in frequency tend to be perceived as coming from the same source. Additionally, the rule of common fate also applies. Sound components that onset and offset together – that is, they have correlated fluctuations in amplitude modulation – tend to group into the same perceptual object. Likewise, grouping by harmonicity can also be phrased in terms of this principle. When a person speaks, their vocal cord vibrations generate energy at a fundamental frequency, as well as at harmonics of this frequency. Thus, the components of their voice can be grouped by identifying acoustic components that have a common spacing in frequency. The principle of closure – referring to the tendency to complete perceptual forms even when evidence is missing – is also thought to be important and applicable to situations when sounds are masked by other sounds. Counter-intuitively, spatial cues seem to only have a weak influence on grouping at this time scale (Darwin & Hukin, 1997; Schwartz, McDermott, & Shinn-Cunningham, 2012) – possibly because they are susceptible to reverberation and interference effects.

Binding over a longer time scale (i.e. stream formation) occurs across acoustic discontinuities where local principles cannot operate and has been posited to depend on the temporal coherence between neural responses that encode various features (Shamma, Elhilali, & Micheyl, 2011). This theory stems from the idea that different sound sources will rarely fluctuate in strength at exactly the same times. The continuity of perceptual features such as frequency (De Sanctis, Ritter, Molholm, Kelly, & Foxe, 2008), timbre (Culling & Darwin, 1993; Cusack & Roberts, 2000), and pitch (Culling & Darwin, 1993; Vliegen, Moore, & Oxenham, 1999) are thought to be important. Spatial cues have also been found to have a stronger effect at this time scale e.g. placing the target sound at a different location from masking sounds as well as setting the target location to be fixed over time are helpful for 'pulling out' a stream (Maddox & Shinn-Cunningham, 2012).

2.4.3 Interactions: Attentional effects on object formation and object-based attention

The two challenges of auditory scene analysis – auditory object formation and selective attention – are closely related processes. However, the exact nature of their interaction remains debated. A particularly controversial issue concerns the role of attention in the formation of streams (Carlyon, Cusack, Foxton, & Robertson, 2001; Cusack, Deeks, Aikman, & Carlyon, 2004; Macken, Tremblay, Houghton, Nicholls, & Jones, 2003; Sussman, Horváth, Winkler, & Orr, 2007). Carlyon et al. (2001) posited that attention is crucial for the effective build-up of streaming. They conducted a study in which participants had to judge the number of streams formed under different conditions and found that the formation of auditory streams in participants is reduced or absent when they attend to a competing task in their contralateral ear. Other studies have nonetheless pointed out potential confounds in their study as well as evidence to support the viewpoint that the initial grouping of acoustic elements can occur pre-attentively (Macken et al., 2003; Sussman et al., 2007). Sussman et al. (2007) conducted a series of mismatch negativity (MMN) experiments using EEG – the MMN is a component of an event-related potential (ERP; see Section 2.5) that is elicited by the brain in response to a deviant stimulus within a sequence – to show that attention is not always required for the formation of streams. Bressler et al. (2014) posited that the truth is likely to lie somewhere in between – both automatic and attention-driven mechanisms interact and play a part in the multi-stage process of forming auditory streams. When low-level attributes are sufficiently distinct, attention is not required to drive the initial segregation of sounds into streams. For more ambiguous mixtures however, it is thought that attention to a particular perceptual feature may help 'pull out' an auditory stream. This is also supported by findings from a study by van Noorden (1975) showing that when subjects listened to sound sequences consisting of two tones with different frequencies, the frequency separation required to induce a percept of two separate streams is smaller if the listener is actively trying to hear out the high-pitch tones than if they are listening less selectively. Additionally, in a cross-modal study (Carlyon et al., 2003) in which subjects listened to tone sequences while performing one of three tasks -- an auditory task related to the stimulus, a visual task or a counting task – it was found that more streaming occurred (i.e. they were more likely to hear two streams) when subjects were performing the auditory task (and therefore were attending to the tones) than when participants performed the other two tasks.

The aforementioned suggestion of how attention and object formation might interact is compatible with the theory of object-based attention. This theory posits that perceptual objects are the basic units on which attention operates, even when attention is first focused on a particular stimulus feature (Alain & Arnott, 2000; Fritz, Elhilali, David, & Shamma, 2007; Shinn-Cunningham, 2008). There is some empirical evidence to support this notion. Best et al. (2008) and Maddox and Shinn-Cunningham (2012) found that discontinuity of other perceptual features of the attended stream, even if irrelevant to the task at hand, can influence task performance, suggesting that the features are bound as an object. Further, when a listener attends to one word, a subsequent word that shares some perceptual feature with the attended word is automatically more likely to be the focus of attention than a word that does not match the preceding word (Bressler et al. 2014). Indeed, this illustrates how attention and stream formation are inextricably linked as it also demonstrates that the continuity of task-irrelevant features of an object/stream can influence the ability to extract information from a sound mixture. Object-based attention also posits that at any given time, the object that is the focus of attention is processed in greater detail than other objects in the scene. Various findings supporting this have now emerged (Chait, de Cheveigné, Poeppel, & Simon, 2010; Nima Mesgarani & Chang, 2012a).

To conclude this section on auditory scene analysis, we revisit the issue concerning the fate of unattended stimuli, a question going back to the early cocktail party studies and filter model theories. It remains debated, but in the last few years, novel neuroimaging analysis techniques have managed to shed some light on this area. Power et al. (2012) found attentional effects on speech processing at the 200-220ms range, suggesting a later locus of selectivity (i.e. post processing of physical characteristics). Puvvada and Simon (2017) found evidence that unattended stimuli are better represented as a single unsegregated object rather than distinct objects in higher-order auditory cortical areas, but that attended and unattended streams are represented with almost equal fidelity in early primary areas. Mesgarani & Chang (2012) conducted an ECoG study and showed that only the attended speaker is represented in STG, which is thought to implement a crucial stage in processing at the transition from acoustic to linguistic processing. Additionally, Brodbeck et al. (2018) and Broderick et al. (2018) found lexical and semantic processing to be categorically selective for attended speech stimuli. These results are consistent with the notion that human auditory cortex is a hierarchical processing system with higher-order processing areas being more affected by attention – for which there has also been evidence from the visual domain (Kastner & Pinsk, 2004; Maunsell & Cook, 2002). Unattended speech streams do not appear to be processed at and beyond the lexical stage when listening to continuous natural speech; a remaining open question is the extent to which pre-lexical processing occurs.

2.5 Electroencephalography (EEG) as a neuroimaging technique

Before the dawn of neuroimaging, scientists relied primarily upon non-human animal studies for understanding lower-level auditory function, and post-mortem examination of the brains of patients with speech and language disorders for studying higher-level speech function. Lesion studies work on the assumption that the observed behavioural deficits in these patients are the result of lesioned brain areas. An obvious disadvantage of this method is that it cannot provide any information about the neural time course of speech processing. But critically, the conclusions that could be drawn about the anatomical localisation of functions were also questionable. Namely, diseased brains do not necessarily function like healthy ones; and these patients often had a variety of pathologies in addition to speech and language issues. The advent of neuroimaging opened up new avenues for studying the cortical processing of speech in healthy brains. fMRI has played an important role in identifying crucial anatomical regions; but temporally-precise methods like EEG/MEG are most suited to the task of understanding the time course of a dynamic cognitive process like speech comprehension. For reasons established in the previous chapter, EEG is the method of choice in this thesis. In this section, a brief overview of the mechanisms behind this method and how it has been classically applied to research is provided.

EEG is a neuroimaging modality for detecting the electrical activity of the brain using electrodes placed on the scalp. It records the summed voltage fluctuations from large neuronal ensembles, reflecting the synchronous post-synaptic activity of a neural population with similar spatial orientation. EEG can record these fluctuations with millisecond temporal resolution, making it an exceptional tool for studying dynamic cortical events. The downside to EEG is that because it is recorded on the scalp, it suffers from volume conduction effects leading to poor spatial resolution (in the order of centimetres). Thus, in general, EEG is not well suited to tackle questions on the precise neuroanatomical localisation of a particular cortical function.

EEG activity contains multiple frequencies simultaneously, which can be separated through signal processing techniques. This activity is conventionally grouped into bands that are defined by logarithmically increasing centre frequencies and frequency widths. Typically, these bands are: delta (2 - 4 Hz), theta (4 – 8 Hz), alpha (8 – 12 Hz), beta (15 – 30 Hz), lower gamma (30 – 80 Hz), and upper gamma (80 – 150 Hz), selected based on neurobiological mechanisms of brain oscillations, including synaptic decay and signal transmission dynamics (Cohen, 2014), although these boundaries are not precise. It has been observed that different cortical functions seem to utilise different frequency ranges. Delta and theta band activity have been linked to speech processing (e.g. Ding & Simon, 2014; Luo & Poeppel, 2007), and laterization of power in the alpha band has been linked to spatial attention (Kerlin, Shahin, & Miller, 2010; Worden, Foxe, Wang, & Simpson, 2000; Wöstmann, Herrmann, Maess, & Obleser, 2016). Studies that employ ECoG often look at signals in the high gamma range, but EEG activity in this range is typically disregarded as it is low pass filtered by the skull and therefore has low signal-to-noise ratio.

Neural activity captured by EEG is traditionally divided into two categories – spontaneous and evoked. Spontaneous activity refers to the neural activity that occurs unprovoked, in the absence of any identifiable stimulus. Recordings of the brain's spontaneous activity is typically used as a diagnostic tool in clinical environments as abnormal states can be detected based on it. It has been found useful for characterising seizures, classifying sleep stages, and monitoring the brain for damage.

Evoked activity, or event-related potentials (ERPs), refers to neural activity generated in response to a specific event. ERPs can be elicited by a wide-range of sensory, motor and cognitive events – e.g. beeps and tones, an oddball stimulus, syntactic violations – and are time-locked to the event in question. As such, they are a useful tool in both diagnostic and experimental research and have been utilised over the years to uncover aspects of the processes that underlie human thought and behaviour. Due to the fact that the EEG reflects thousands of simultaneously ongoing processes, ERPs are not usually visible in single-trial EEG data. The typical method for acquiring a visible ERP is to average epochs of EEG responses time-locked to a particular discrete stimulus event across multiple repetitions – the idea here is that the background, non-phase-locked activity should average out to near zero given a sufficient number of repetitions, leaving only the stimulus-related activity that is consistent across trials.

Averaged ERP waveforms consist of a series of positive and negative voltage deflections, which are categorised as individual components. A component is a scalp-recorded voltage change that reflects a specific neural or psychological process (Luck et al., 2012). The latency and amplitude of components are used to make inferences regarding the underlying psychological processes as well as the likely site of origin. Traditionally, researchers have found it useful to group components into two classes. Components that depend on the physical properties of sensory stimuli are termed 'exogenous' components, whilst components whose characteristics vary as a function of higher-level factors as attention and task relevance are called 'endogenous' components. Nonetheless, strict categorisation as one or the other has been proven to be an oversimplification. ERP components that occur within the first 100ms of stimulus presentation tend to be more exogenous, while those occurring later tend to be more endogenous. But early components have also been shown to be modifiable by cognitive manipulations (e.g. attention), and many of the later 'cognitive' components have been shown to be influenced by the physical attributes of the eliciting condition (Coles & Rugg, 1995).

Components are typically referred to by a letter indicating their polarity (N for negative and P for positive), followed by a number indicating their latency in milliseconds. When components have the same polarity and similar latencies, scalp distribution is then used to distinguish between them.

ERP experimental paradigms have been utilised extensively by neuroscientists over the years to test how an external manipulation influences a specific component of the elicited ERPs. Depending on the stimulus modality, the ERP will typically have a series of early components, reflective of initial processing in a sensory receiving area, followed by ones that reflect more integrative cognitive processing. An ERP evoked by an auditory stimulus is termed the Auditory Evoked Potential (AEP). The AEP includes components spanning the full length of the auditory pathway (an example is shown in Figure 2-7).

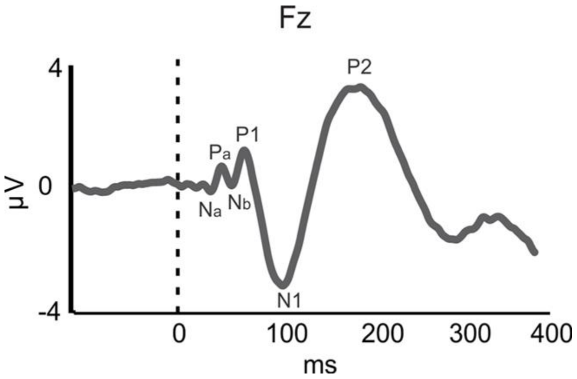


Figure 2-8 | The Auditory Evoked Potential (AEP)

An averaged AEP waveform in response to a piano tone at the frontal channel, Fz (Shahin, 2011).

Research on AEPs in response to speech suggest that components with longer latencies relate to progressively higher levels of the speech processing hierarchy (Salmelin, 2007). Specifically, early responses come from the brainstem, the middle-latency responses derive from an initial activation of the auditory cortex, and late responses come from auditory and associated cortices (Picton, 2011). Several language-comprehension-related components have been identified at long latencies – the N400 and P600 are generally associated with semantic and syntactic violations in sentences (although results contradicting this hypothesis have been found, so the precise underlying causes of these components remain debated; see Kuperberg (2007) for review).

2.6 Encoding/Decoding methods and the Temporal Response Function (TRF) framework

As established in the previous section, traditional ERP studies require stimuli to be simple (only varying across the to-be-tested dimension), discrete, and presented to subjects multiple times so that a time-locked averaged response potential can be computed. While such studies have been useful for helping us identify relevant stimulus features, their findings do not necessarily generalise to continuous natural stimuli like speech (Hamilton & Huth, 2018). For one, the experimental paradigms used are artificial and unlike scenarios that would be encountered in the real-world. The last few years have seen a rise in the employment of encoding/decoding methods to sensory neuroscience research. These methods overcome the aforementioned limitations of the ERP technique, allowing for the use of continuous stimuli that vary along multiple dimensions, thus, enabling researchers to conduct experiments using complex stimuli such as natural speech. Encoding entails modelling the activity of a neural signal as a function of stimulus features, while decoding refers to using neural activity to predict/reconstruct a stimulus feature (Holdgraf et al., 2017).

One framework for implementing encoding with EEG/MEG signals involves making the simplifying assumption that the brain is a linear, time-invariant (LTI) system. The mapping between stimulus features and the (time-lagged) neural signal is computed using regularised linear regression and the resulting estimated system response is termed the temporal response function (TRF) (Crosse, et al., 2016; Lalor & Foxe, 2010). We refer to this framework as the TRF approach. The LTI assumption is not strictly true but has been found to be a reasonable one under certain circumstances (Bialek et al., 1991; Lalor and Foxe, 2010; Ding and Simon, 2012; Mesgarani and Chang, 2012; Pasley et al., 2012). With respect to the auditory domain, the TRF approach has been shown to be effective for estimating a temporally-precise electrophysiological response to uninterrupted natural speech (Lalor & Foxe, 2010).

2.6.1 System response (TRF) estimation

Within the TRF framework, the stimulus property, $s(t)$, is taken to be the input to the system (i.e. the brain), and the recorded N-channel EEG signal response, $r(t,n)$, to be the output. The response model can then be represented as (Crosse et al., 2016):

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

Here, $\varepsilon(\mathrm{t},\mathrm{n})$ is the residual response at each channel *n* and time point *t* not explained by the model; $w(\tau,n)$ is the impulse response or TRF that describes the linear transformation of the ongoing stimulus to the ongoing neural response for a specific range of time lags, τ, relative to the instantaneous occurrence of the stimulus feature, $s(t)$. This time lag window is included to account for the fact that speech processing in the brain is not instantaneous and occurs over a period of time.

The TRF is estimated by minimising the mean-squared error (MSE) between the actual and estimated neural responses. This minimisation problem can be solved using reverse correlation, expressed in the following equation:

$$w = (S^T S)^{-1} S^T r$$

Here, r is the matrix of EEG data with channels arranged column-wise, and S is the lagged time series of the stimulus feature $s(t)$:

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

where τ_{min} and τ_{max} are the minimum and maximum time lags (in samples) respectively. Each time lag in S is arranged column-wise and non-zero lags are padded with zeros to ensure causality (Mesgarani, David, Fritz, & Shamma, 2009).

In practice, to avoid any ill-posed estimation problems and prevent overfitting, the TRF approach adds a regularization term to the autocovariance matrix, $S^T S$, prior to its inversion (Crosse et al., 2016):

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

Here, I is the identity matrix, and λ is the ridge parameter, or bias term, which is determined using cross-validation. This particular form of regularization is known as Tikhonov regularization, or ridge regression (Tikhonov & Arsenin, 1977).

The above model of the system assumes a univariate input signal. Nonetheless, it is easily adaptable for the simultaneous testing of multiple feature dimensions. For an input signal of Z variables, the stimulus lag matrix is simply extended such that every column is replaced with Z columns (essentially, this is akin to computing a lag matrix for each variable, then concatenating them to form a single matrix). The resulting TRF model can be unwrapped accordingly to acquire the weights corresponding to each individual variable.

2.6.2 Decoding/Stimulus reconstruction

Stimulus-response mapping can be performed in both directions. Encoding, as described above, models neural activity as a function of uni- or multivariate stimulus features. The resultant TRF characterises the system, revealing the spatio-temporal dynamics of the neural response as indexed by EEG to the input speech parameter. Decoding (also known as backward modelling or stimulus reconstruction) on the other hand derives reverse stimulus-response mappings in order to predict stimulus features from the neural response. Decoding exploits all of the available EEG data simultaneously to derive the stimulus-response transformations. Thus, this method is more sensitive to small differences between EEG channels that are highly correlated with each other, as is often the case with EEG. Estimating the decoder model is carried out in much the same way as in the case of encoding, except with the stimulus and EEG response in place of each other in all the above equations, and with the time lags in reverse direction as the model effectively maps backwards in time (Crosse et al., 2016). To reverse the lags, the values of τ_{min} and τ_{max} are swapped but their signs remain unchanged. In the case of decoding, we refer to the estimated system response as the 'decoder'. Unlike the TRF, the decoder is not easily interpretable in a neurophysiological sense (Haufe et al., 2014); nonetheless, the weights reflect the channels that contribute the most towards reconstructing the stimulus feature.

2.6.3 Fitting the TRF/decoder and evaluating model performance

Fitting the TRF/decoder entails selecting a ridge parameter (i.e. lambda value, λ) that will optimise stimulus-response mapping. As mentioned above, this is done using leave-one-out cross-validation. That is, for a given data set with *M* trials, individual system responses (TRF/decoder) are computed using a wide range of lambda values (e.g. 1, 1x10^2, … 1x10^9) for all but one trial (i.e. M-1 trials). For each lambda value, an average TRF/decoder is attained by averaging over the *M*-1 TRF/decoders. The set of average models are then used to predict the EEG response/stimulus feature of the left-out trial. Model performance is assessed by quantifying how accurately the predicted EEG response/reconstructed stimulus feature matches the original response/feature. This metric often used is the Pearson's correlation coefficient and the resulting coefficient is referred to as prediction accuracy of the model. The entire procedure is then repeated *M*-1 times, rotating the trial that is tested such that each trial is 'left-out' of the training set once. The overall model performance for each lambda value can then finally be determined by averaging over the individual model performances for each trial. The lambda value that results in the best prediction/reconstruction accuracy is selected (Crosse et al., 2016).

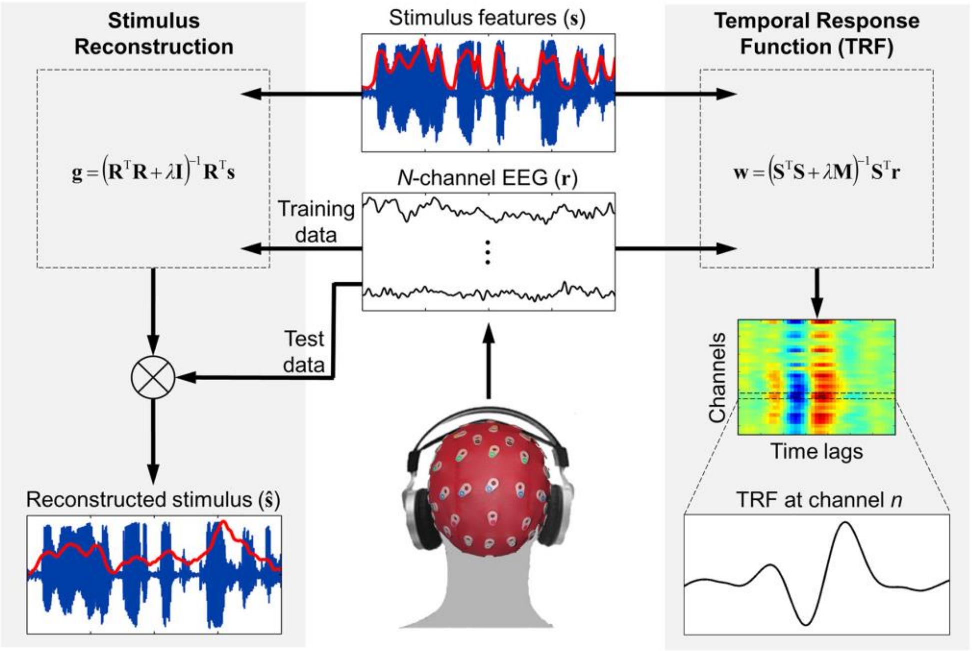

Stimulus Reconstruction
Stimulus features (s)
Temporal Response Function (TRF)
$\mathbf{g} = (\mathbf{R}^T\mathbf{R} + \lambda\mathbf{I})^{-1}\mathbf{R}^T\mathbf{s}$
N-channel EEG (r)
$\mathbf{w} = (\mathbf{S}^T\mathbf{S} + \lambda\mathbf{M})^{-1}\mathbf{S}^T\mathbf{r}$
Training data
Test data
Channels
Time lags
Reconstructed stimulus (ŝ)
TRF at channel n

Figure 2-9 | The TRF framework

Summary of the TRF framework for stimulus-response mapping (Crosse et al., 2016). Encoding (as shown on the right) yields a system response (i.e. the TRF) that can be used to predict EEG data from the stimulus. A prediction accuracy that is significantly above chance reflects EEG tracking of the stimulus feature. The TRF is also interpretable – it reveals the spatio-temporal dynamics of the neural response to the input stimulus feature. Decoding (or stimulus reconstruction; shown on the left) can be used to reconstruct the stimulus feature from EEG data. As with the case of encoding, the accuracy of this reconstruction is indication of neural tracking of the stimulus feature.

2.6.4 Dependent measures produced by the TRF framework

The TRF framework is summarised in Figure 2-8. Because the TRF approach allows for the use of continuous stimuli that vary across multiple dimensions, it has been highly effective for studying the processing of natural speech. Employing this framework to representations of speech based on linguistic theories has enabled the investigation of how the brain represents speech through the cortical hierarchy, as well as how the various processing stages are affected by attention. Thus far, the TRF approach has been successful at isolating EEG/MEG neural indices at various levels.

The TRF approach was first shown to be effective for modelling the neural response to the acoustic envelope of speech. The acoustic envelope – slow temporal fluctuations in the overall signal amplitude between the rates of about 2 to 50Hz (Rosen, 1992) – contains essential cues that are necessary for the understanding of speech. It conveys various acoustic and linguistic information – e.g. segmental cues for phonetic features, and prosodic cues. A breakthrough finding in cognitive neuroscience in recent years has been the discovery that low-frequency MEG (Ahissar et al., 2001; Luo and Poeppel, 2007) and EEG (Aiken and Picton, 2008) activity is phase-locked to the acoustic envelope. As such, it was a natural choice for testing the efficacy of the TRF framework. It was found that a temporally-detailed TRF could be obtained when relating the speech envelope to EEG (Lalor & Foxe, 2010). Additionally, when stimulus reconstruction was performed, the speech envelope could also be reconstructed from EEG activity at above chance (Crosse et al., 2016). It has since also been demonstrated (though testing subjects under different listening conditions) that higher-level cognitive functions such as selective attention and intelligibility can modulate the accuracy of this reconstruction (Ding and Simon, 2012; Mesgarani and Chang, 2012; Zion Golumbic et al., 2013; O'Sullivan et al., 2015; Vanthornhout et al., 2018). Indeed, its sensitivity to attention has been shown to be useful for decoding attention in the context of the cocktail party paradigm (O’Sullivan et al., 2015). O’Sullivan et al (2015) trained a decoder mapping from EEG to the envelope of the attended speech stream. They found that for a new trial, the signal reconstructed by the decoder correlates stronger with the envelope of the attended speech stream than with that of the unattended speech stream. By exploiting this finding, attentional selection could be decoded with an average accuracy of 89% for ~60-second-long single-trial EEG data. This is one of the state-of-the-art methods for decoding auditory attention.

Because the envelope tracking effect has been found to be so robust, it has been a major focus of EEG/MEG speech research in the last few years. However, as mentioned earlier, it is a univariate aggregate measure that conveys many acoustic and linguistic cues, so it is insufficient for disambiguating and characterising the precise computations involved in the processing of natural speech. Thus, there have been recent efforts to relate EEG/MEG activity to other representations of speech that more precisely characterise particular processing stages. Di Liberto et al. (2015) recorded high-density scalp EEG from subjects as they listened to continuous natural speech and employed the TRF approach to show that low-frequency EEG is best modelled when representing speech in terms of acoustics plus phoneme/phonetic feature labels. Brodbeck et al. (2018) modelled speech in terms of a variety of representations reflecting lexical-level processing. These representations were based on the premise that phonetic information is used to incrementally constrain possibilities for the word that is currently being processed. In relating the features to corresponding MEG data, they found TRFs indicative of phoneme-level predictive coding and lexical competition. Broderick et al. (2018) used a computational model to quantify the meaning contained in words based on how semantically dissimilar they were to their preceding context. They found that the resultant TRF displayed a prominent negativity at a time lag of 200-600 ms on centro-parietal EEG channels, akin to the N400 ERP component. Notably, this semantic signature was absent when subjects were not attending to the speech. Together, these studies demonstrate the efficacy of the TRF method for probing the computations involved at various levels of the speech processing hierarchy within the context of more naturalistic scenarios.

2.7 Summary

In this chapter, we reviewed the literature concerning the brain basis of speech comprehension, focussing on the pathways from sound to meaning and the role of selective attention. We found gaps in the knowledge within several subareas. Namely, the debates concerning early versus late selection (i.e. the fate of unattended speech), and whether a phoneme-level mapping exists in speech comprehension remain unresolved. We also discussed EEG and encoding/decoding methods, focussing on the TRF approach that will be implemented in this thesis. We found that this approach has been highly effective for isolating neural signatures of acoustic and linguistic speech representations that are computed during natural speech processing. Nonetheless, there remain representations essential for understanding speech which have not been explored – the processing of prosodic information being one. Lastly, we found that reconstruction of the acoustic envelope is state-of-the-art for auditory attention decoding. But given that there are other features also modulated by attention, we hypothesise that looking beyond using envelope reconstruction alone to devise a joint approach may lead to improved decoding accuracy.

Chapter 3 Decoding the target talker in a non-stationary cocktail party scenario

3.1 Introduction

As established in the previous chapters, a real-life cocktail party environment poses a difficult listening challenge – there are a handful of conversations happening simultaneously around the room, and the interlocutors are often non-stationary. For the hearing-impaired, it is a listening situation that is particularly effortful. However, the auditory systems of most normal-hearing humans are remarkably adept at dealing with such complex scenes, affording them the ability to attend to a single talker and even track them as they move.

The precise computations that occur in the human brain to perform this feat has long been a topic of interest. We reviewed the growing body of auditory scene analysis research examining the mechanisms behind this phenomenon in Section 2.4. Of late, there has been increasing interest in how this research can be applied to the development of next generation, cognitively-controlled, smart hearing aids (Lunner et al. 2009; Mirkovic et al., 2016). Studies in this vein are concerned with the task of decoding selective auditory attention from non-invasive neural measurements, the end goal being to track in real-time and in a real-life setting the talker to which one is attending.

In this context, the findings that cortical activity tracks the temporal envelope of speech (Ahissar et al, 2001; Luo et al, 2007; Aiken et al., 2008), and that there is a separate and enhanced representation of the attended speaker in neural activity as indexed by magneto-/electroencephalography EEG/MEG (Ding & Simon, 2012; Nima Mesgarani & Chang, 2012b) offer a promising avenue. By relating the amplitude envelope of a speech stimulus to simultaneously recorded MEG/EEG activity (Ding and Simon, 2012; O'Sullivan et al., 2015; Horton et al., 2014), it has been shown that single trial data can be decoded offline to ascertain attentional selection. In these studies, a cocktail party scenario was simulated using simple diotic or dichotic paradigms (with no spatial filtering) in which subjects were cued to attend to one of two competing talkers. Building upon these works, recent efforts have looked at the practicalities of implementing this decoding approach in daily life (Mirkovic et al., 2015), its efficacy in a more realistic binaural listening environment (Das et al., 2016), and optimising the decoding process towards real-time application (Akram et al., 2016; 2017). The effect of real-world background noise on decoding performance (Das et al., 2018) has also been considered. Additionally, there have been efforts towards integrating noise suppression and auditory attention decoding into one system (Das et al., 2017; O'Sullivan et al., 2017; Van Eyndhoven et al., 2017).

In this chapter, we describe a paradigm for investigating auditory attention decoding in an environment in which the to-be-attended talker is non-stationary – a common scenario in real life. We test the effectiveness of the envelope reconstruction (i.e. backward modelling) decoding framework established in O'Sullivan et al. (2015) in this setting. We then consider the incorporation of other signatures of attention to improve decoding and track talker location. Namely, we hypothesise that the computed 'decoders' themselves could be exploited to improve decoding. Decoder values reflect the weightings applied to the multichannel EEG data at different time lags and different scalp locations that optimally reconstruct the amplitude envelope of the stimulus. Although not easily interpretable from a neurophysiological standpoint, decoders are more robust than forward (stimulus-to-EEG) models as they utilize the multivariate nature of the recorded neural signal. We will show that the decoders display different spatio-temporal patterns as a function of a subject's attended target speaker. We also explore using the topographic distribution of power in the alpha band, as it has been observed in previous studies to be hemispherically lateralised in accordance with the deployment of attention to the right or left side of space (Frey et al., 2014; Haegens et al., 2011; Wöstmann et al., 2016). The results of this study were published in the *Journal of Neural Engineering* (Teoh & Lalor, 2019).

3.2 Methods

3.2.1 Subjects

Nine female and five male subjects between the ages of 19 and 30 took part in the study. All subjects were right handed and spoke English as their primary language. Subjects were of normal hearing – they self-reported no history of hearing impairment or neurological disorder. Each subject provided written informed consent prior to testing and received monetary reimbursement. The study was approved by the Research Subjects Review Board at the University of Rochester.

3.2.2 Stimuli and Procedures

Subjects undertook 10 trials, each of 90 seconds in length. Stimuli consisted of two Sherlock Holmes stories narrated by a female and a male talker. Silent gaps in the audio exceeding 0.3 s were truncated to 0.3 s in duration.

The two speech streams were filtered using Head-Related Transfer Functions (HRTFs), giving rise to the perception that the talkers were at 90 degrees to the left and right of the subject. HRTFs are The HRTFs used were obtained from the CIPIC database (Algazi et al., 2001) – the same HRTFs were used for all participants. To simulate a dynamic environment with two non-stationary interlocutors (or rapid head turns by the listener), the talkers were instantaneously alternated between the left and right locations at short but varying intervals, ranging between 4 – 22 seconds (Figure 3-1).

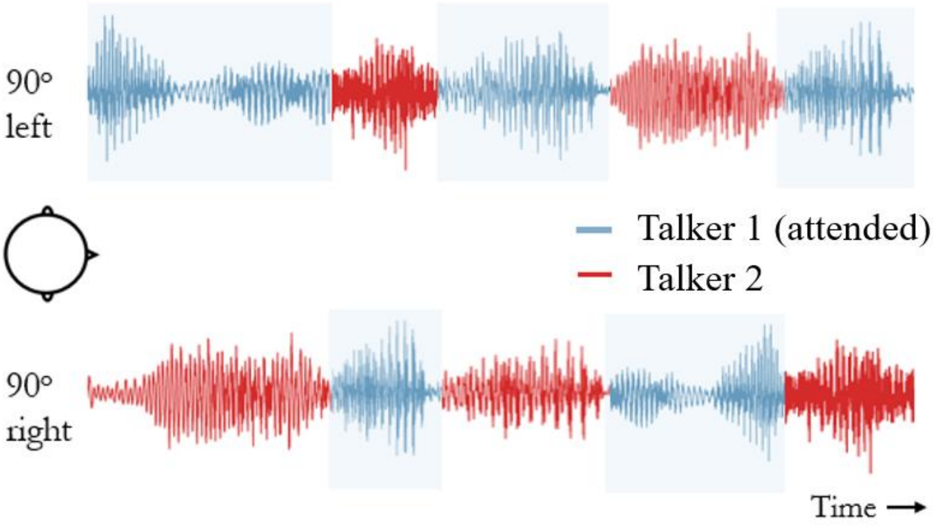


Figure 3-1 | Stimuli for the non-stationary talker paradigm.

Subjects listened to two concurrent talkers. To simulate non-stationarity, the two talkers were alternated between the left and right locations at short but varying intervals.

Subjects were instructed to selectively attend to the story narrated by the male talker. They were asked to minimize motor activities and to maintain visual fixation on a crosshair centred on the screen during trials. After each trial, subjects were required to answer four multiple-choice comprehension questions (each with four choices) on both the attended and unattended stories.

Subjects also undertook a control condition (with a larger number of trials but only the first 10 were used here so as to match the moving condition) in which the talkers were stationary throughout each trial, replicating the paradigm employed in O'Sullivan et al (2015), albeit with HRTF-filtered stimuli rather than dichotic.

All stimuli were sampled with a frequency of 44 100 kHz and were presented using Sennheiser HD650 headphones and Presentation software from Neurobehavioral Systems (http://www.neurobs.com).

3.2.3 Data Acquisition and Pre-processing

The experiment was conducted in a soundproof room. A Biosemi ActiveTwo system was used to record EEG data from 128 electrode positions (digitized at 512Hz).

EEG data were referenced to the average of all scalp channels. Automatic bad channel rejection and interpolation was performed. A channel was deemed bad if the standard deviation of the channel was lower than a third of or exceeded three times the mean of the standard deviation of all channels. In place of the bad channel, data were interpolated from the four nearest neighbouring electrodes using spherical spline interpolation (Delorme & Makeig, 2004). To decrease subsequent processing time, data were down-sampled to 128Hz. Data were filtered into two frequency pass bands: (1) 1-8 Hz for decoder computation and envelope reconstruction, selected based on earlier studies (O'Sullivan et al., 2015; Das et al., 2016), and (2) 8-14 Hz for computation of alpha band power (see below).

Amplitude envelopes of the original audio stimuli and the left and right audio streams simulating non-stationary talkers were computed. This was done by taking the absolute value of the Hilbert transform of the stimuli, applying a 30Hz low-pass filter, and down-sampling to 128Hz.

3.2.4 TRF implementation

Like in O'Sullivan et al. (2015), the TRF implementation of backwards modelling (i.e. using regularized linear regression) was employed to relate the multivariate neural data to the attended amplitude envelopes.

We described TRF estimation in Section 2.6. Essentially, we denote a multivariate spatio-temporal filter/decoder $g(\tau, n)$ that linearly maps from the EEG response back to the stimulus envelope. The decoder g is estimated by minimizing the MSE between the actual and estimated envelopes, and is computed using the following matrix operation:

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

where λ is the ridge regression parameter, I is the identity matrix, and the matrix R is the lagged time series of the response matrix, r.

The mTRF Toolbox (Crosse et al., 2016 - https://sourceforge.net/projects/aespa/) was used to solve for the decoders. All decoders in this study were computed for the time lag interval of 0 to 250 ms. This reflects the idea that changes in the amplitude envelope of the ongoing stimulus are likely to produce effects in the ongoing EEG at time lags of 0 to 250 ms, an idea that is supported by previous research (Lalor & Foxe, 2010). The ridge regression parameter was determined using leave-one-out cross-validation, selected to maximise the average Pearson's correlation coefficient between actual and predicted stimuli.

3.2.5. Decoding Attentional Selection

3.2.5.1 Effects of Attended Talker Movement on Decoding Accuracy

To evaluate whether decoding as assessed by envelope tracking is robust to switches in the location of the attended talker, we compared single-trial decoding accuracies of our non-stationary talker paradigm to that of our control (stationary talker) condition. Specifically, for both conditions, we implemented a previously introduced analysis (O'Sullivan et al., 2015) as follows (Figure 3-2): Attended decoders that mapped from the 1-8 Hz EEG data to the speech envelope of the talker to which they were instructed to attend were computed for each subject and each trial. For a particular subject, we could then reconstruct the stimulus envelope of a particular trial, *N*, using the average attended decoder of *N-1* trials (i.e. leave-one-trial-out cross validation). The Pearson's correlation coefficient, *r*, was computed between the reconstructed envelope and both the actual attended and unattended stimulus envelopes. A trial was deemed correctly classified if the reconstructed envelope was more correlated with the attended than the unattended envelope ($r_{attended} > r_{unattended}$).

We also tested for switching costs over a shorter time scale by extracting 20-second long epochs in which there were two, three and four switches in attended talker location (moving talker condition), as well as no switches (stationary talker condition). Ten epochs were extracted for each condition.

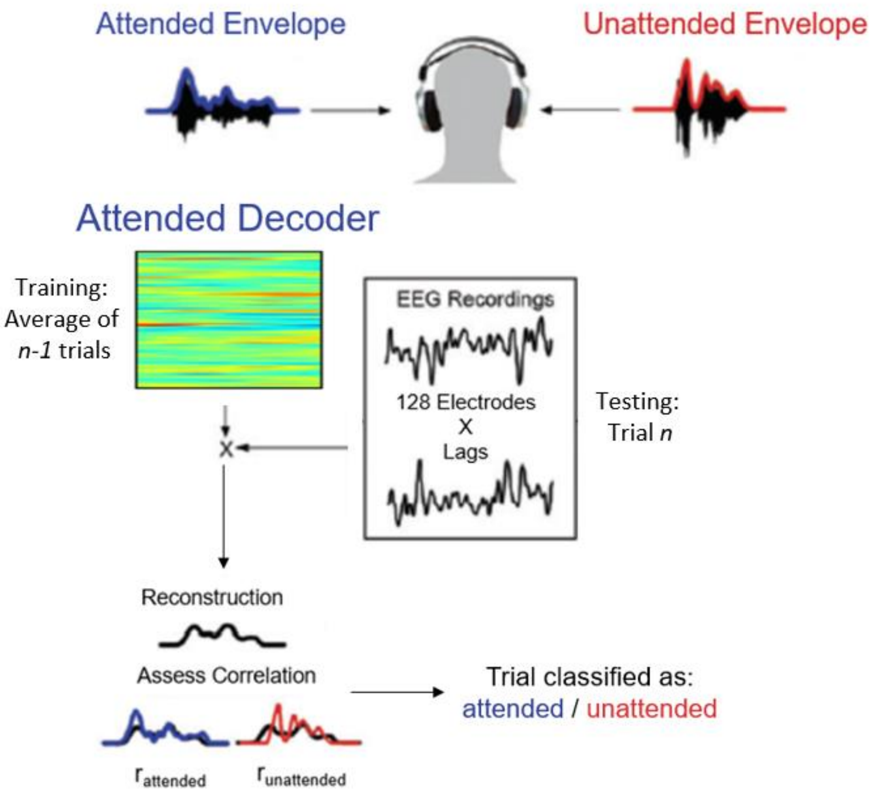


Figure 3-2 | Illustration of the decoding strategy (adapted from O'Sullivan et al., 2015). *An estimate of the attended speech envelope was obtained from each segment of neural data. Decoders were trained using n-1 trials and averaged, then tested on the nth trial. The correlation between the reconstructed stimulus and the attended and unattended speech envelope were assessed, resulting in reconstruction accuracies $r_{attended}$ and $r_{unattended}$. If the reconstructed envelope was more correlated with the attended than the unattended envelope ($r_{attended} > r_{unattended}$), the trial was considered to be correctly classified. This was repeated n times, rotating the trial to be tested each time (leave-one-out cross-validation).*

3.2.5.2 Spatio-temporal Decoder Structure

To assess if the decoder structure itself – that is, the weightings applied to the different EEG electrode channels at different time lags – contained spatio-temporal signatures of attention that were consistent across subjects and could be exploited, we extracted epochs from the 1-8 Hz EEG data such that the talkers were stationary within each epoch. Epochs were between 4 – 22 s in length. For each epoch, we then computed decoders mapping from the EEG data to the corresponding left and right audio envelope segments. A fixed regularisation parameter – selected to optimise envelope reconstruction accuracy over all epochs – was used for training. Supposing that the deployment of attention (A vs U) and the spatial locations of the talkers (L vs R) are reflected in the decoder values, the following two combinations of decoder types would arise: if a subject is attending to the left during a particular segment, the decoder mapping from their EEG to the left audio envelope will be a ‘left attended’ decoder (A_L) and the decoder mapping to the right audio envelope will be a ‘right unattended’ decoder (U_R); conversely, if a subject is attending to the right, the decoder mapping from their EEG to the left audio envelope will be a ‘left unattended’ decoder (U_L) and the decoder mapping to the right audio envelope will be a ‘right attended’ decoder (A_R).

We tested whether we could exploit the consistency of decoder weight patterns across subjects to decode the attended talker during each epoch. We did this in a leave-one-subject-out cross-validation manner. For *N-1* subjects, we concatenated segments of data and stimuli based on whether they were attending to the right or left, and computed the four aforementioned decoder types (A_L, U_R, U_L and A_R). We then averaged these decoders across subjects and concatenated them to form two exemplar decoder weight pattern matrices ([A_L, U_R] and [U_L, A_R]). For each epoch of the left-out subject, we calculated the Pearson’s correlation between a concatenated matrix of decoders mapping from EEG to the left and right envelopes with these matrices (i.e., with [A_L, U_R] or [U_L, A_R]). A classification was deemed correct if the tested decoder matrix was more correlated with the grand-average decoder of its actual category than that of the opposing class.

The previously described analysis was based on using the decoder weights on all 128 channels at all time lags between 0 and 250 ms. In addition to this, we also wanted to explore if there were particular decoder weights or time lags that might be most informative for decoding the attended talker (this was partly motivated by the longer-term goal of conducting this type of analysis on a much reduced, wearable EEG system, e.g., Mirkovic et al., 2016). To do this, we tested if decoding could be improved by selecting features based on a group-level cluster-based permutation test of the effect across all time points and scalp electrodes. Specifically, we computed single-subject averages for each decoder type then carried out non-parametric cluster-based tests (Maris & Oostenveld, 2007; Oostenveld et al., 2011) to compare attended and unattended decoders (A_R vs U_R and A_L vs U_L). The tests revealed spatio-temporal regions in which differences were most pronounced.

As an aside, we also repeated the classification and feature selection steps to investigate whether there were significant differences between attended decoders (A_L vs A_R); that is, if the *direction* of attention itself was discernible from the attended decoders.

3.2.5.3 Alpha Band Power

Previous research has shown hemispheric lateralisation effects in alpha band power during spatial attention tasks (Frey et al., 2014; Haegens et al., 2011; Wöstmann et al., 2016; Bednar et al., 2018). To test whether that could be useful in this setting for decoding cocktail party attention, we extracted 4-seconds of 8-14 Hz EEG data time-locked to the onset of each switch (4-seconds corresponds to the length of the minimum switch interval). The absolute value of the Hilbert transform of each epoch was computed as a measure of alpha band power. We then tested if the epochs could be classified based on the attended direction (‘attend left’ versus ‘attend right’). To do so, we calculated the Pearson’s correlation between the spatio-temporal distribution of alpha power for each epoch of each subject (normalised by the sum of mean ‘attend right’ and mean ‘attend left’ epochs) with the spatio-temporal distribution of alpha power averaged over all ‘attend left’ and ‘attend right’ trials (also normalised) for all the other subjects. A classification was deemed correct if the tested decoder was more correlated with the grand-average decoder of its actual category than that of the opposing class.

Again with a view towards identifying which channels and time lags might be most informative, we followed up with an exploratory analysis where we tested if decoding could be improved by selecting features based on a group-level cluster-based permutation test of the effect. Single-subject averages were computed and non-parametric cluster-based tests (Maris & Oostenveld, 2007; Oostenveld et al., 2011) were carried out to compare normalised ‘attend left’ and ‘attend right’ epochs. Only spatio-temporal regions in which differences were most pronounced were used for decoding.

3.2.5.4 Tracking the Locus of Attention

The previous analyses were focused on treating each epoch where the talkers were stationary as a stand-alone trial. However, we also wondered if we could continuously track subjects' locus of attention even as the attended talker switched locations. To do this, we carried out an overlapping sliding window analysis. We chose a window size of 4-seconds (which corresponds to the minimum duration between changes in talker location) and a step size of 31.25ms. For each window, the envelope reconstruction decoding analysis described in section 3.2.5.1 was performed and reconstruction accuracies were assessed.

As with the main analysis above, we then looked to include spatio-temporal decoder information to see if it improved the decoding for these shorter segments. For each window, we computed the decoders mapping the EEG data to both the right and left audio stream envelopes. We evaluated the correlation of the spatio-temporal decoder structure with trained grand average attended and unattended decoders (left and right) (computed in the same manner as in section 3.2.5.2, i.e. using data from all other subjects).

A support vector machine (SVM) classifier was trained with the correlation coefficients as features in a leave-one-trial-out manner, and then tested on each window of the remaining trial. To determine whether incorporating spatio-temporal signatures improved decoding, classification was first carried out using only the correlation coefficients from the envelope reconstruction analysis, and then a combination of those features with the correlation coefficients from the decoder structure analysis.

In all our analyses, chance level was established using non-parametric permutation testing, where the null distribution was determined by permuting labels for trials 1000 times. The tail of the empirical distribution was then used to calculate the p-value for the original classification.

3.3 Results

3.3.1 No significant difference in decoding accuracy due switching of talker location

Subjects' responses to the comprehension questions for both the moving and stationary (control) conditions revealed that they were compliant. Percentages of correct answers for the attended story (means = 75.8±3 % and 78.4±3 % for the moving and stationary conditions, respectively) significantly exceeded those of the unattended story in both cases (two-tailed Wilcoxon signed-rank; z= 3.3; p = 0.00097 for both conditions), where accuracies for the unattended story were close to the chance level of 25% (means = 29.4±2.5 % and 25.9±2 % for moving and stationary, respectively). There was no significant difference between moving and stationary conditions (two-tailed Wilcoxon signed-rank; z=-1.18; p=0.237). The lower average for the moving talker condition could be attributed to participants performing relatively poorly in the first trial (mean = 48.2±9.3 %) as they adapted to the task. The average performance in the moving talker condition for trials after the first was similar to the stationary condition (mean = 78.9±2.8 %).

A comparison of envelope reconstruction decoding accuracies revealed a higher mean average performance for the moving talker condition (mean = 85.7±3.7 %, as opposed to 77.9±4.2 % for the stationary condition) (Figure 3-3A), however this difference was not found to be significant (two-tailed Wilcoxon signed-rank; z =1.78; p=0.075). We also found no significant differences in decoding accuracy between shorter (20 s) epochs with two switches, and those with three and four switches within the moving talker condition (Figure 3-3B) (two-tailed Wilcoxon signed-rank; z= -0.9078; p= 0.3640, and z= 0.2379; p= 0.8119). This was also the case when those decoding accuracies were compared to the stationary talker condition (two-tailed Wilcoxon signed-rank; two switches vs stationary - z=-1.19; p=0.2317; three vs stationary - z=-0.43; p=0.6684; four vs stationary – z =-0.898; p=0.3689).

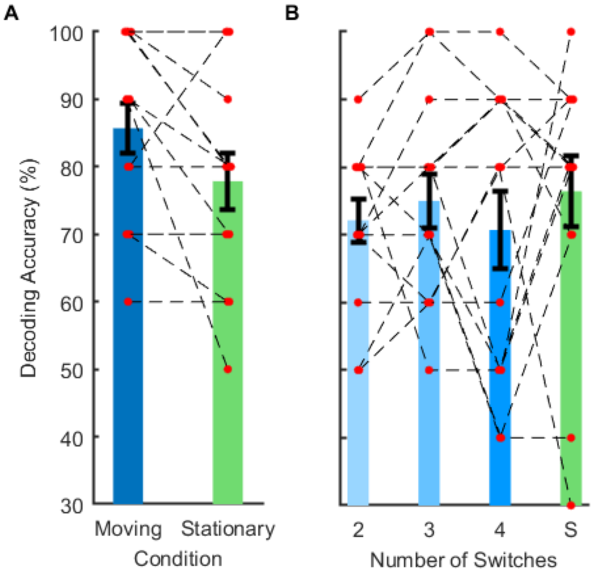


Figure 3-3 | Decoding accuracies as a function of switching in talker location

(A) Decoding accuracies for the moving and stationary talker conditions were on average 85.7% and 77.9% respectively. This difference was not found to be significant. (B) Decoding accuracies for epochs with two, three, and four switches within the moving talker condition, and for epochs with no switches taken from the stationary talker condition. No significant difference was found within the moving talker condition and between the moving and stationary conditions. Error bars represent mean ± standard error.

3.3.2 Attentional selection is reflected in spatio-temporal decoder structure

EEG data (1-8Hz) were epoched into segments such that the talkers were stationary within each epoch. Training and classification of epochs was performed as described in Section 3.2.5.2. We found that it was possible to significantly classify the talker to which subjects were attending within single epochs based on the decoder weight patterns themselves (Figure 3-4A All Features, mean = 70.7 ± 2.08%; two-sided permutation test; p=0.001).

A group-level cluster-based permutation test (two-sided dependent samples t-test, minimum of one neighbourhood channel, critical alpha-level of 0.05, Monte Carlo method for calculating significance probability) of the effect across all time points and scalp electrodes revealed significant differences between attended and unattended decoders (A_L vs U_L and A_R vs U_R). These differences were most pronounced over large spatiotemporal regions of the decoder structure, particularly between around 50-110 ms and 140-210 ms (shown at a representative latency within these clusters – 70 ms and 190 ms – in Figure 3-4C).

Having identified spatio-temporal regions at which differences between attended and unattended decoders were most pronounced, classification was repeated using only these features (i.e. clusters that were significant at the alpha-level of 0.05). This however did not result in improved average decoding accuracy (Figure 3-4A Selected Features, mean = 69.52 ± 2.10%) and the difference between using selected features and all features was found to be insignificant (two-tailed Wilcoxon signed-rank test; z= -1.69; p= 0.09).

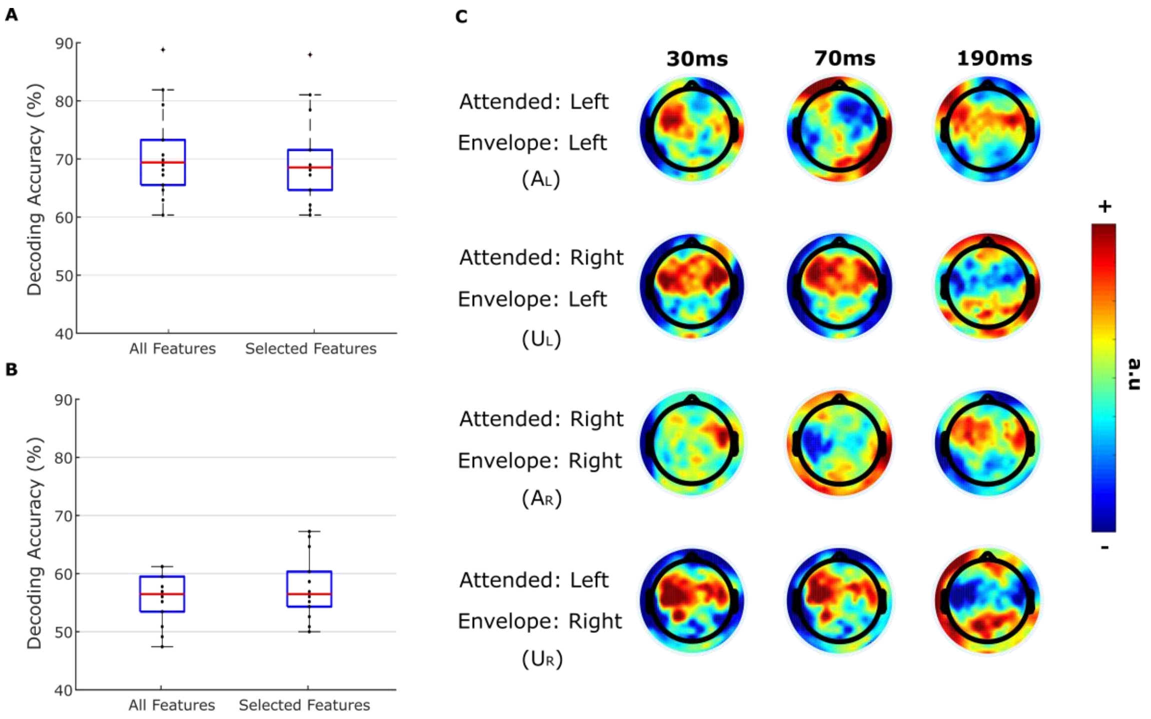


Figure 3-4 | Decoding attentional selection and direction using decoder model weights

(A) Accuracies of decoding of attentional state by classifying decoder weight patterns ([A_L, U_R] vs [U_L, A_R]). Decoding accuracies when all features are used and when only selected features are used are depicted. Black dots represent individual subject data. (B) Classification of 'attend right' vs 'attend left' decoders (A_R vs A_L) with and without feature selection. (C) Grand average decoder weights at representative latencies at which differences between decoder types were most pronounced based on a cluster-based permutation test. The first and second rows depict attended and unattended decoders which map from EEG data to left-sided stimuli, while the third and fourth rows depict attended and unattended decoders mapping EEG data to right-sided stimuli.

It was also possible to differentiate attended decoders ('attend right', A_R vs 'attend left', A_L) at above chance level (Figure 3-4B All Features, mean = 55.3 ± 1.11%; two-tailed permutation test; p=0.001). A group-level cluster-based permutation test in this case revealed a particularly pronounced difference between ~25-80 ms in left central sensors (shown at a representative latency, 30 ms, in Figure 3-4C). Classification using only these features resulted in an increased average decoding accuracy (Figure 3-4B Selected Features, mean = 56.53 ± 1.21%) but this difference was not found to be significant (two-tailed Wilcoxon signed-rank test; z=0.52; p=0.599).

3.3.3 Alpha power lateralisation is insufficiently robust for decoding attentional selection

A group-level cluster-based permutation test of the effect across all time points and scalp electrodes found no significant difference in alpha band power between ‘attend left’ and ‘attend right’ epochs (two-sided dependent samples t-test, minimum of one neighbourhood channel, critical alpha-level of 0.05, Monte Carlo method for calculating significance probability). A subsequent test on the signals averaged across the full 4-second time window also revealed no significant effect. We considered the possibility that alpha band power lateralization occurs when there is a switch in the location of the attended/ignored talker but is not consistently maintained throughout the trial (Kerlin et al., 2010). Thus, we averaged the signals over non-overlapping 100ms time windows and tested (again, using a group-level cluster-based permutation test) for differences within each window. With this test at a critical alpha-level of 0.05, the differences between the ‘attend left’ and ‘attend right’ conditions were significant for the latency ranges of 500-700ms, 1400-1500ms, 3400-3600ms, and 3900-4000ms post-switch. This must be considered as exploratory and read with caution given that these effects were only found after averaging within certain windows. For what it is worth, the normalised contrast between grand average alpha band power when subjects were attending to the left and right (‘attend left’ minus ‘attend right’, over ‘attend left’ plus ‘attend right’) at these latencies are shown in Figure 3-5A.

It was possible to classify the unaveraged signals based on subjects’ direction of attention at above chance level only after feature selection was carried out (All features: mean = 51.3±1.08%; two-sided permutation test; p=0.126; Selected features: mean = 55.6±1.21%; two-sided permutation test; p=0.001). That is, for each 100ms time window, only channels in which the effect was pronounced within the window were included (Figure 3-5B). As such, while it is known that alpha power lateralizes with spatial attention, the effect does not seem to be strong enough to produce robust signatures that can be used for decoding attention to a speech stream on a single trial level, at least in the current experiment.

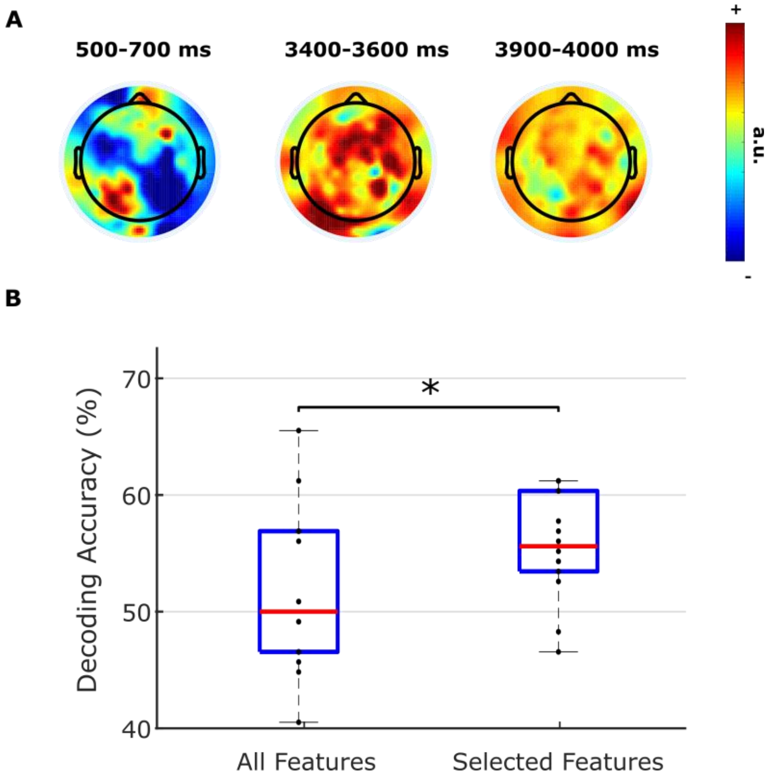


Figure 3-5 | Decoding attentional selection using alpha power

(A) Normalised contrast, also known as the Attention Modulation Index (Wöstmann et al., 2015) between grand average alpha band power (that is, 'attend left' minus 'attend right', divided by 'attend left' plus 'attend right') shown at latency ranges in which the difference between conditions was found to be significant (B) Box plots showing accuracies of decoding of attentional direction using alpha band power with and without feature selection. Individual subject data is denoted by the black dots. A significant improvement was found (two-tailed Wilcoxon signed rank test; z= 2.15; p=0.03) after which direction of attention could be classified, but only after deliberately searching for and selecting specific spatio-temporal features, suggesting that alpha lateralization is not a robust enough signal to reliably help with decoding attention in the context of the current experiment.

3.3.4 Tracking the locus of attention

An overlapping sliding window analysis was carried out to determine if we could continuously track the talker to which subjects were attending. This was implemented by using an SVM classifier to decode attentional selection for each window based on envelope reconstruction correlation values (*env*), decoder weight correlations (*DW*), as well as a combination of the two (*env+DW*) (Figure 3-6B). All approaches led to decoding accuracies that were significantly greater than chance (two-tailed permutation test; p=0.001 in all three cases). We also found a significant increase in decoding accuracy when spatio-temporal decoder weights were included along with the envelope reconstructions (*env* – mean = 61.07±2.27%; *DW* – mean = 61.76±1.11%; *env+DW* – mean = 65.31±1.57%) (two-tailed Wilcoxon signed-rank test; *env+DW* vs *env*– z=3.23; p=0.0012; *env+DW* vs *DW*– z=3.23; p=0.0012).

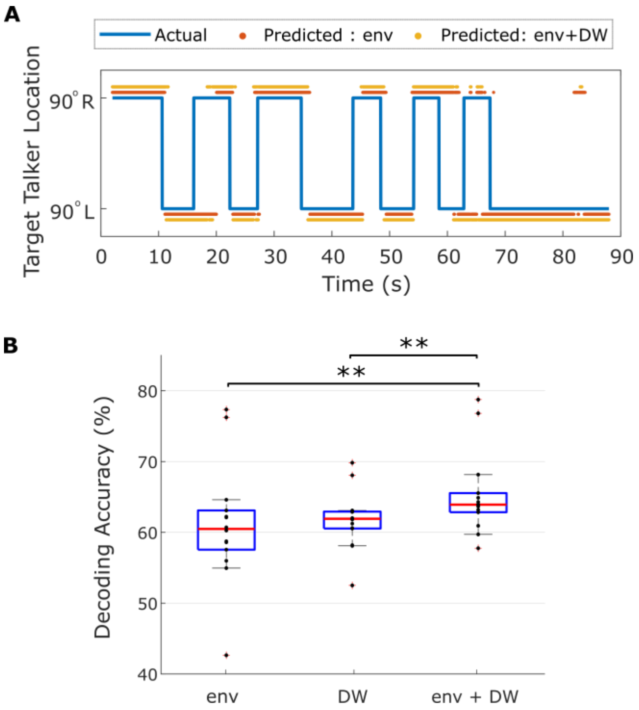


Figure 3-6 | Tracking the locus of attention

*(A) Actual and predicted attended talker location over time for one trial of one example subject. Predictions are based on envelope reconstruction (env) as well a combination of envelope reconstruction and decoder weights (env+DW). (B) Boxplots of decoding accuracy based on a sliding 4 s window of EEG data when implementing stimulus envelope reconstruction alone (env), spatio-temporal decoder weights (DW) and when using a combination of the two ('Env+DW') (**p<0.01). The black dots represent data from individual subjects.*

3.4 Discussion

This study looked at auditory attention decoding in an environment in which the to-be-attended talker is non-stationary. We set out to test the efficacy of the envelope reconstruction approach in this scenario, as well as investigate whether the spatiotemporal structure of our decoders and alpha band power lateralization could be utilised to improve decoding.

Our motivation for testing decoder performance in a non-stationary scenario stems from the findings of previous auditory scene analysis studies. In Section 2.4, we reviewed auditory object formation, selective attention, and how the two processes are intertwined. A prominent idea in the field is that attention acts on perceptual objects, but the formation of an object over time depends on the continuity of all stimulus perceptual features (Alain & Arnott, 2000; Shinn-Cunningham, 2008). Further, when attention is sustained on an auditory object, perception of the object improves over time. Consistent with this idea, previous studies involving discontinuity of features in the to-be-attended stream have found costs in behavioural task performance. Best *et al* (2008) conducted a spatial attention study in which subjects were tasked with attending to one of two simultaneous but spatially-separated audio streams of rapidly-presented discrete digits under different target voice conditions (i.e. same target talker from digit-to-digit versus different target talker). Stimuli were presented via loudspeakers and subjects were cued visually towards a target location by LEDs affixed to the loudspeakers. They found that when spatial location and target voice were both held constant (Experiment 2 in their paper), there was an enhancement in digit recall performance compared to conditions where location was fixed but target voice was varied (Experiment 1). When they introduced spatial location switches into both experiments, the switches brought about a reduction in performance accuracy within both experiments, but this reduction was larger within Experiment 2. Maddox and Shinn-Cunningham (2012) and Bressler et al. (2014) also found that listeners were more error-prone when a sound feature, even if task-irrelevant, changed and broke down the perceived continuity of the stream.

In our study, subjects were not asked to attend to a particular feature or cued towards a location prior to/at a switch. They were merely informed that the to-be-attended talker would be male and would at first instance be on their right but would then continually swap locations with the other (female) talker. Given the lack of explicit spatial cueing, it is possible that the strategy employed here by subjects to guide their deployment of attention would be to use vocal (e.g. pitch and timbre) or perhaps even high-level semantic information instead of spatial features. Nonetheless, generalisation of the aforementioned findings would suggest in our control (stationary) condition when both talker voice and spatial location are constant, there would be an enhancement over the moving talker condition due to sustained attention to an auditory object. We did not find evidence for this behaviourally – there was no overall significant difference between how subjects performed in the moving condition in comparison with the control condition. This could be due to the nature of the task which perhaps lacked the sensitivity to pick up on any differences in performance. The stimuli used were continuous narratives and subjects had to answer comprehension questions based on content rather than recall individual words. Subjects may have still suffered a cost, but it is possible that the cost was so transient that it did not impact too strongly on what was a very high-level and relatively coarse measure of attention. Nevertheless, subjects did perform significantly poorer in the first trial of the moving talker block compared to subsequent trials and the stationary block, suggesting that they found the task with location switches more difficult and required time to adapt – perhaps adjusting how they deployed attention.

In terms of EEG attentional decoding, our results demonstrate that the decoding approach established in O'Sullivan *et al* (2015) is robust to switches in the attended talker's location. In fact, decoding accuracies for 7 subjects were better for the moving condition compared to the stationary condition, whilst it was only better for one subject in the stationary condition compared to the moving condition. The overall average decoding accuracy was found to be higher in the moving condition, although this was not significant. It is conceivable that subjects had to expend more effort to perform the task in the moving talker condition, leading to a higher signal-to-noise ratio in their EEG. This in turn could have compensated for any costs due to the switches. Indeed, it has previously been observed that the inclusion of moderate background noise results in an improvement in decoding performance over a clean speech condition (Das, Bertrand, & Francart, 2018).

As a further check whether there is a significant change to decoding accuracy when the attended talker moves, 20 second epochs containing varying number of switches (2, 3 and 4) from within the moving talker condition were extracted. No significant differences in decoding accuracies were found between the 3 cases. Additionally, we also cut 20 second epochs from the stationary condition for comparison. Again, no significant differences were found between the stationary and moving conditions, but here the overall average was slightly higher in the stationary condition.

Although backward models are not easily physiologically interpretable (Haufe et al., 2014), signatures of attentional selection and attended direction were found to be present in the spatio-temporal structure of decoders and were consistent across participants. Indeed, we could exploit the consistency of the encoded attentional selection information across subjects to significantly improve decoding accuracy over using envelope reconstruction alone when continuously tracking subjects' locus of attention with short sliding windows. Admittedly, the analysis was carried out with the inclusion of EEG channels across the whole scalp, and it could be that this improvement is contingent on this. Future work would include reducing the number of channels used to test its practicality for implementation in a wearable device. Signatures of attended direction, which we tested for by comparing 'attend left' and 'attend right' decoders were weaker but could still be decoded at above chance level on a single-trial basis.

Recent spatial attention studies have found that alpha band power is hemispherically lateralized according to the direction of attended stimulus (Kerlin et al., 2010; Horton et al., 2014; Frey et al., 2014; Wöstmann et al., 2015; van Diepen et al., 2016; Bednar et al., 2018). In particular, studies have reported that alpha is observed in the preparatory/anticipatory period (if present for the task) and between 300-700ms after stimulus onset. Despite our task not being a spatial attention task per se (and indeed it is likely that subjects guided their attention based on vocal features or story content) the talkers were always spatially separated. So, we were interested to see if alpha band lateralization was present and could be exploited to detect switches in location. We found that the lateralization was present but weak and not maintained steadily throughout the trial – on a group-level, it was only found to be significant within certain latency ranges after a switch, and only after searching through many latencies. In particular, it was observed at ~0.5-0.7 s after the onset of each switch – this is consistent with previous studies. The significant lateralization observed at the end of the trial (~3.4 and ~3.9 seconds) may be due to anticipation of the next switch in location, although caution must be taken to not overinterpret this result. One possible explanation for the overall weak alpha lateralisation effect is that spatial location is not the specific feature attended to here. Indeed, the presence of alpha lateralisation at all here may be because auditory attention is directed towards perceptual objects so bounded spatial features are also enhanced, albeit to a lesser degree.

Using a stationary talker cocktail party paradigm, Horton *et al* (2014) had previously found decoding using alpha band power lateralization to be ineffective. They reasoned that it could likely be because their analysis window only began 1000ms after stimulus onset, later than the frequently observed lateralisation peak at ~0.6 s. Here, our window included the 0.6 – 0.7 s range; nonetheless, decoding accuracy remained significantly lower than when using envelope and approximately similar to their results. Feature selection improved decoding accuracy of the epochs such that all but one subject could be decoded at above chance. However, the weakness of the effect and the fact that it appears to not be maintained throughout the trial precludes the usefulness of its incorporation into a continuous decoder for improving current state-of-the-art decoding methods, at least based on the current experimental paradigm.

3.5 Summary

In this chapter, we tested the efficacy of a previously established envelope reconstruction framework for decoding attention within a non-stationary talker scenario. We found no significant difference in decoding performance even when the attended talker switched their location using this approach. We then showed that the consistency of spatio-temporal decoder weights across participants could be exploited to improve decoding accuracy within the context of our experiment. This suggests that looking beyond envelope reconstruction to incorporate other metrics into the framework has potential to improving current decoding approaches. Lastly, we considered alpha band power lateralization as another possible feature for decoding selective attention. It was found to reflect the spatial location of the attended talker at certain latencies, but the strength of this effect was not robust enough to produce reliable single-trial decoding.

Chapter 4 Isolating indices of prosodic pitch processing in low-frequency EEG

4.1 Introduction

Speech is arguably humanity's most important signal because of its extraordinary power to convey information. The structure of the speech signal and the strata of information that it contains was reviewed in Section 2.3. We established that meaning in speech is largely conveyed by the selection of linguistic units such as words, phrases, and sentence; but that prosody, the thread that ties these segments together, also adds another layer of meaning. Prosody is made up of the suprasegmental variations of acoustic cues – the lengthening of a syllable, the rise and fall of pitch across a phrase, the increase in loudness at a word. These rhythms and melodies of speech enable the same sequence of words to convey different meanings and emotions (Bolinger, 1986).

Because of prosody's importance to spoken language, it is likely that we have evolved specific neurophysiological mechanisms to process this additional channel of information. Many studies have sought to anatomically localise prosodic processing in the brain (Baum & Pell, 1999; Gandour et al., 2004; Kreitewolf et al., 2014; Meyer et al., 2002; Plante et al., 2002; Sammler et al., 2015; Witteman et al., 2011). But another fundamental question concerns *how* prosody is encoded. The quest to uncover its neural correlates is complicated by the fact that systems for characterising prosodic elements in speech are less established (Cole & Shattuck-Hufnagel, 2016).

Nonetheless, as mentioned in Section 2.3.3, one of the most important perceptual cues underlying prosody is pitch – its modulation giving rise to phenomena such as intonation and stress. Pitch can be estimated from the acoustic speech stream using algorithms such as autocorrelation (Boersma, 1993). A few recent studies have investigated the encoding of pitch in cortex. In an fMRI study, it was shown that pitch-sensitive regions respond primarily to resolved harmonics (Norman-Haignere et al., 2013). And, in an ECoG study specifically investigating speech prosody, Tang *et al*. (2017) found that high-gamma activity on a number of single electrodes over superior temporal gyrus selectively represented intonation contours in terms of relative pitch.

In this chapter, we investigate the encoding of pitch during listening to continuous narrative speech in ongoing scalp-recorded electroencephalography (EEG). Like ECoG, EEG has high temporal resolution but with markedly less sensitivity to high-gamma activity. However, as mentioned previously, EEG has the advantages of being portable and non-invasive, meaning that it can be used to study speech and language processing in a wider variety of subjects. It is well-established that low-frequency EEG tracks the temporal speech envelope and it was recently shown that using the spectrogram to represent acoustic energy leads to improved EEG prediction (Di Liberto et al., 2015). But disambiguating responses to other features related to the neural processing of speech can be difficult as most features covary and are nested in the envelope. Isolable neural indices at the level of phonemic (Di Liberto et al., 2015) and semantic processing (Broderick et al., 2018) have recently been found, but it is unknown if pitch processing can also be dissociated.

We employ forward modelling to regress EEG on two pitch-related measures – harmonic resolvability (Norman-Haignere et al., 2013), on which pitch saliency is dependent, and relative pitch (Tang et al., 2017). We test if these features significantly predict EEG after accounting for features previously shown to index acoustic/phonetic processing and examine their response characteristics. Given that different EEG frequency bands have been linked to different roles in speech processing (Giraud & Poeppel, 2012), we also examine if these indices may be more evident in particular bands. The findings from this chapter have been published in the *European Journal of Neuroscience* (Teoh et al., 2019).

4.2 Methods

4.2.1 Subjects

All subjects spoke English as their primary language and had no reported history of hearing impairment or neurological disorder. Our study involved analysing data from two experiments. Nineteen subjects (13 male) took part in the first experiment and thirteen subjects (8 male) took part in the second. Both experiments were approved by the Ethics Committee of the School of Psychology at Trinity College Dublin. Each subject provided written informed consent. These data have been previously analysed using different methods and published (Experiment 1 - Broderick et al., 2018; Di Liberto et al., 2015; Experiment 2 - Di Liberto, Crosse, & Lalor, 2018). Data for the first experiment can be found at http://datadryad.org/resource/doi:10.5061/dryad.070jc.

4.2.2 Stimuli and Procedures

In the first experiment, subjects undertook 20 trials, each of between 2 to 3 minutes in length, in which they were presented with an audiobook version of a classic work of fiction ('The Old Man and the Sea' by Ernest Hemmingway) read by a professional male American English narrator. Subjects were asked to simply listen to the story (there was no behavioral task), and they were allowed to take breaks in between trials.

The second experiment was originally designed to test the effects of prior knowledge on the encoding of natural speech (Di Liberto, Crosse, & Lalor, 2018). We utilized data from a segment of the experiment to validate our pitch-related measures. In this particular segment, subjects were presented with a 10-second snippet of noise-vocoded (Davis & Johnsrude, 2003; Shannon et al., 1995) speech, followed by a clean speech snippet. The noise-vocoded and clean speech snippets have highly correlated acoustic envelopes, but whilst clean speech possesses clear pitch, vocoded speech does not. The snippets were randomly selected from the same audiobook as the first experiment. In the majority of trials ('standard' trials), the noise-vocoded speech snippet was a degraded version of the subsequently-presented clean speech snippet (specifically, the clean speech was filtered into three frequency bands and the amplitude envelope of each band was used to modulate band-limited noise). However, in some of the trials, the vocoded and clean speech snippets did not match ('deviant' trials). Subjects were tasked with identifying if a trial was standard or deviant after listening to the clean speech. We only analyzed data from standard trials (78 ten-second trials in total per subject).

All stimuli were presented monophonically at a sampling rate of 44100 Hz using Sennheiser HD650 headphones and Presentation software from Neurobehavioral Systems (http://www.neurobs.com). Testing was carried out in a dark room and subjects were instructed to maintain visual fixation for the duration of each trial on a crosshair centered on the screen, and to minimize eye blinking and all other motor activities.

4.2.3 Data Acquisition and Pre-processing

A Biosemi ActiveTwo system was used to record EEG data from 128 electrode positions digitized at 512 Hz in both experiments. All channels were referenced to the average of two mastoid channels and downsampled to 128 Hz. EEG data were filtered between 0.2 to 30 Hz (broadband), and into delta (0.2 – 4 Hz), theta (4 – 8 Hz), alpha (8 – 15 Hz), and beta (15 – 30 Hz) bands using Butterworth filters. Instantaneous power and phase were computed for the narrowband filtered data by taking the absolute value and the cosine of the angle of the Hilbert transform of the signals respectively.

We represent our speech stimuli in terms of two pitch-related parameters – pitch (f_{rel}), and harmonic resolvability (*res*) (example waveforms of the pitch measures, as well as the temporal envelope for comparison, are shown in Figure 4-1). f_{rel} quantifies pitch, normalised according to the vocal range of the speaker, and harmonic resolvability (*res*) is a measure related to whether the harmonics of a sound can be processed within distinct filters of the cochlea (resolved) or if they interact within the same filter (unresolved). It was important to ensure that any neural tracking of these pitch measures was not explainable in terms of the various acoustic and phonetic features that have previously been shown to predict neural responses (Di Liberto et al., 2015). As such, we included a number of other speech representations in our analysis framework, including acoustic energy measures – the temporal envelope (*env*) and spectrogram (*s*) – as well as phonetic features (*f*). The measures were computed as follows:

Pitch-related measures -

- f_{rel}: Praat software (Boersma & Weenink, n.d.) was used to extract a continuous measure of pitch (fundamental frequency/absolute pitch) at a sampling rate of 128 Hz. It performs pitch estimation using an autocorrelation method (Boersma, 1993). The z-score of this measure was then computed to obtain speaker-normalised relative pitch on a per-speaker basis
- *res*: Following fMRI work on this topic (Norman-Haignere et al., 2013), we extracted a measure of harmonic resolvability based on a model of the human auditory periphery (McDermott & Simoncelli, 2011)

Acoustic/phonetic representations (collectively, *efs*) -

- *env*: The stimuli were first filtered between 80 Hz and 2800 Hz. The absolute value of the Hilbert transform of the resulting signal was then taken. The envelopes were further low-pass filtered at 30 Hz and downsampled to 128 Hz
- *s*: Stimuli were filtered into 16 frequency bands between 250 Hz and 8 kHz according to Greenwood's equation (Greenwood, 1961), and then the amplitude envelope for each frequency band was computed by taking the absolute value of the Hilbert transform
- *f*: Prosodylab-Aligner software (Gorman, Howell, & Wagner, 2011) was used to perform forced-alignment, generating phoneme-level (American English International Phonetic Alphabet; IPA) segmentation of the stimuli. Each phoneme was then mapped to a set of 19 features based on their manner of articulation (University of Iowa's phonetics project; http://prosodylab.cs.mcgill.ca/tools/aligner/). Based on the start and end times of each feature, a multivariate time-series binary matrix was produced to mark when each feature was present in the stimuli

The Pearson's correlation between f_{rel} and *res* with the envelope were found to be 0.187 and 0.843 respectively.

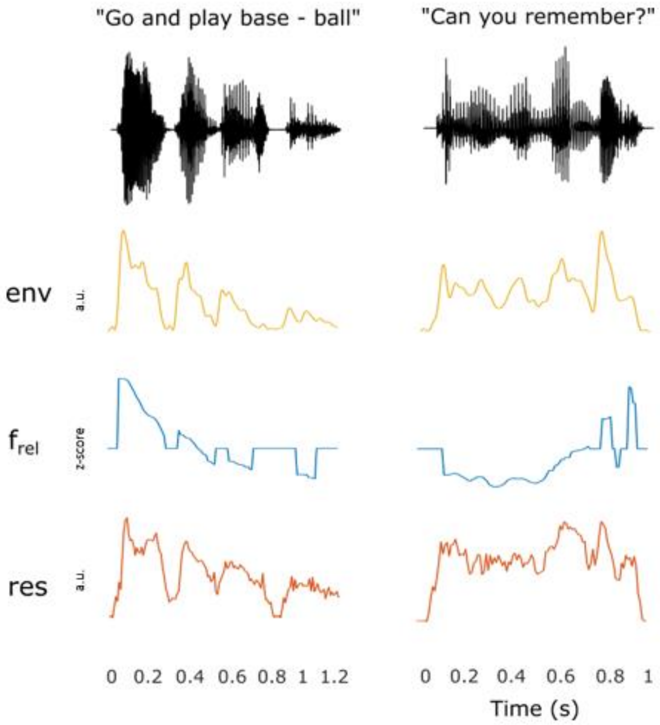


Figure 4-10 | Stimuli examples

Subjects listened to an audiobook version of a classic work of fiction read by a professional narrator. Pitch-related measures – relative pitch (f_{rel}) and harmonic resolvabiliy (res) – and the temporal envelope estimated from the audio signal of example statement ('Go and play baseball') and question ('Can you remember?') phrases from the book are depicted. Statements in English generally end with a falling intonation contour, whilst questions end with a rising contour as can be seen in the f_{rel} measure.

4.2.4 TRF implementation

The different speech representations described above were mapped to the concurrently-recorded 128-channel EEG signals. Multivariate regularized linear regression was employed to relate the speech features to the recorded EEG data (i.e. encoding), where each EEG channel is estimated to be a linear transformation of the speech features over a range of time lags. This transformation is described by the TRF, detailed in Section 2.6.

We use the mTRF toolbox (Crosse et al., 2016 - https://sourceforge.net/projects/aespa/) to solve for the TRF using reverse correlation with ridge regression. The ridge regression parameter was tuned using leave-one-out cross validation. That is, we trained on *n*-1 trials for a wide range of λ values (e.g. 1x10^1, 1x10^2, … 1x10^9), computed the average TRF across trials for each λ, then tested the TRFs on the *n*th trial. This was repeated *n* times, rotating the trial to be tested each time. The λ value that maximized the Pearson's correlation coefficient between the actual and predicted neural response over all trials was selected.

The transformation was computed over time lag intervals of 0 to 300ms. This was selected based on previous speech-related studies, where no visible response was present outside this range (Di Liberto, O'Sullivan, & Lalor, 2015; Lalor & Foxe, 2010). Using a time lag interval reflects the idea that changes in the features of the ongoing stimulus are likely to produce effects in the ongoing EEG in that interval. We quantified how well each speech representation related to the neural data using leave-one-trial-out cross-validation (as described above), with Pearson's correlation coefficient as our metric of prediction accuracy. Because the cross-validation procedure takes the average of the validation metric across trials, the models are not biased toward the test data used for cross-validation (Crosse et al., 2016).

To evaluate whether a pitch feature contributed independently of all other features in predicting the neural data, we computed the partial correlation coefficients (Pearson's r) between the EEG predicted by each pitch measure's model with the actual recorded EEG after controlling for the effects of all other acoustic/phonetic and pitch features that were found to predict EEG at above chance level.

4.2.5 Statistical Analysis

To test that an individual feature predicts EEG at above chance level, we performed non-parametric permutation testing. The neural responses were permuted across trials such that they were matched to features from a different trial, and the same leave-one-out cross-validation procedure as described above was performed to compute TRFs and prediction accuracies. This was done 1000 times for each subject to establish a distribution of chance-level prediction accuracies. Based on these prediction accuracies, we also computed a distribution of partial correlations for the pitch-related measures (i.e. we partialled out the contributions of all other features).

To perform group-level statistical testing, we generated a null distribution of group means. One prediction accuracy from each subject's individual distribution was selected at random to go into each group mean. This process was repeated 1000 times, sampling with replacement for each subject.

4.3 Results

4.3.1 Pitch-related features uniquely contribute to predictions of broadband EEG

We regressed the individual pitch and acoustic/phonetic representations to broadband EEG (0.2 – 30 Hz). Figure 4-2 shows the prediction accuracies for the pitch-related measures (averaged over the best 30 channels for each feature; but using all channels did not alter the pattern of results), along with their topographic distributions. We found that all features could individually predict EEG better than chance (one-tailed permutation test, f_{rel}: p=9.99e-4, *res*: 9.99e-4, *env*: 9.99e-4, *f*: 9.99e-4, *s*: 9.99e-4). Of course, there is some redundancy in these features – there is overlap between the acoustic/phonetic measures, and as mentioned above, we found correlation between the pitch features and envelope. Given that the acoustic/phonetic features are tracked by cortex, it is therefore not surprising to also find significant tracking of the two pitch-related measures. One way to examine whether they contribute any unique predictive power is to test for improvement in prediction accuracy when a pitch-related feature is added to the joint acoustic/phonetic model (*env*+*f*+*s*, or *efs* for short). We found this to be the case – both pitch-related measures improved EEG prediction above and beyond that of the joint acoustic/phonetic model (Wilcoxon Signed Rank test (vs *efs*) – f_{rel}+*efs*: z=3.299, p=9.6733e-04, *res*+*efs*: z=2.213, p=0.0269, f_{rel}+*res*+*efs*: z=3.702, p= 2.1367e-04). The results of these combined models are also depicted in Figure 4-2.

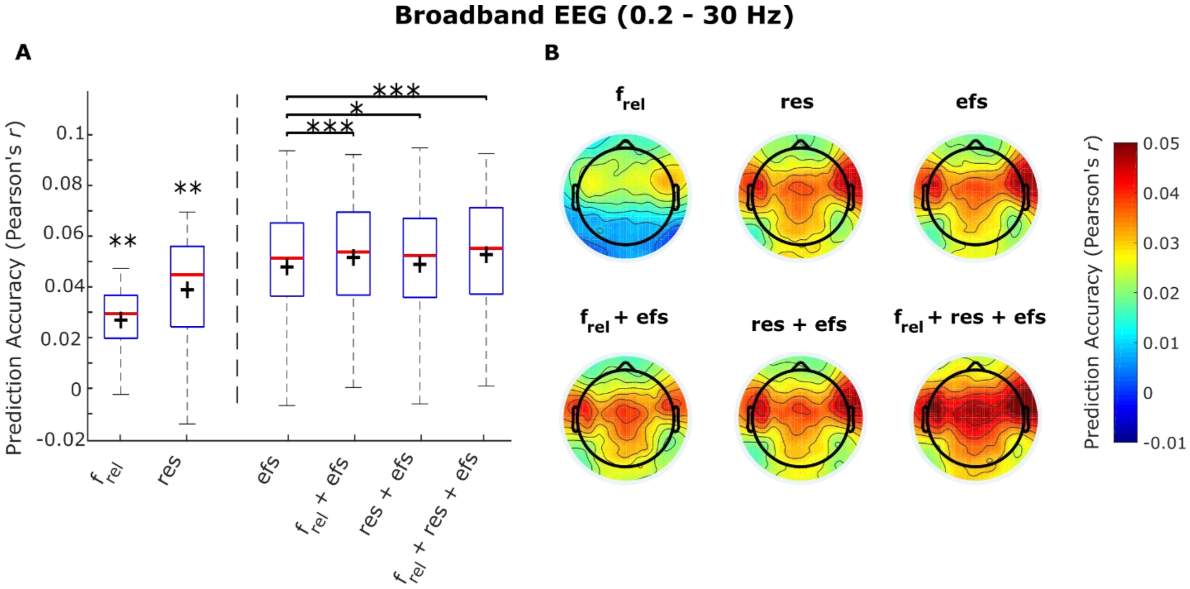


Figure 4-2 | Broadband EEG results

*Broadband EEG prediction accuracies (Pearson's r) for the two pitch-related measures (f_{rel} & res; averaged across the best 30 channels for each representation/combination although using all channels did not alter the significance of results) are shown on the left side of plot (A). To determine if these features have unique predictive power, we check if their inclusion improve prediction accuracies above and beyond the joint acoustic/phonetic model (efs). The prediction accuracies of these combined models are shown on the right side of plot (A). The black crosses indicate mean across all subjects and asterisks indicate statistical significance (for individual measures: one-tailed permutation test; for paired comparison of the combined models to efs: Wilcoxon signed-rank test; *p<0.05, **p<0.01, ***p<0.001). Topographic distribution of prediction accuracies across all subjects are shown in (B).*

A second way of checking for unique predictive power is to compute the partial correlations of the EEG predicted by the pitch-related features with the actual EEG after controlling for all other features (including the other pitch representation). An advantage of this method is that the specific contribution of each measure can be more clearly characterized. Figure 4-3 depicts the partial correlations of the pitch features and their topographic distributions. Non-parametric permutation testing revealed significant partial correlations (on a group-level) for both measures (one-tailed permutation test, f_{rel}: p=9.99e-4, and *res*: p=0.0020). On an individual subject level, we found significant partial correlation for 14 of the 19 subjects for f_{rel}, and 4 of the 19 subjects for *res*.

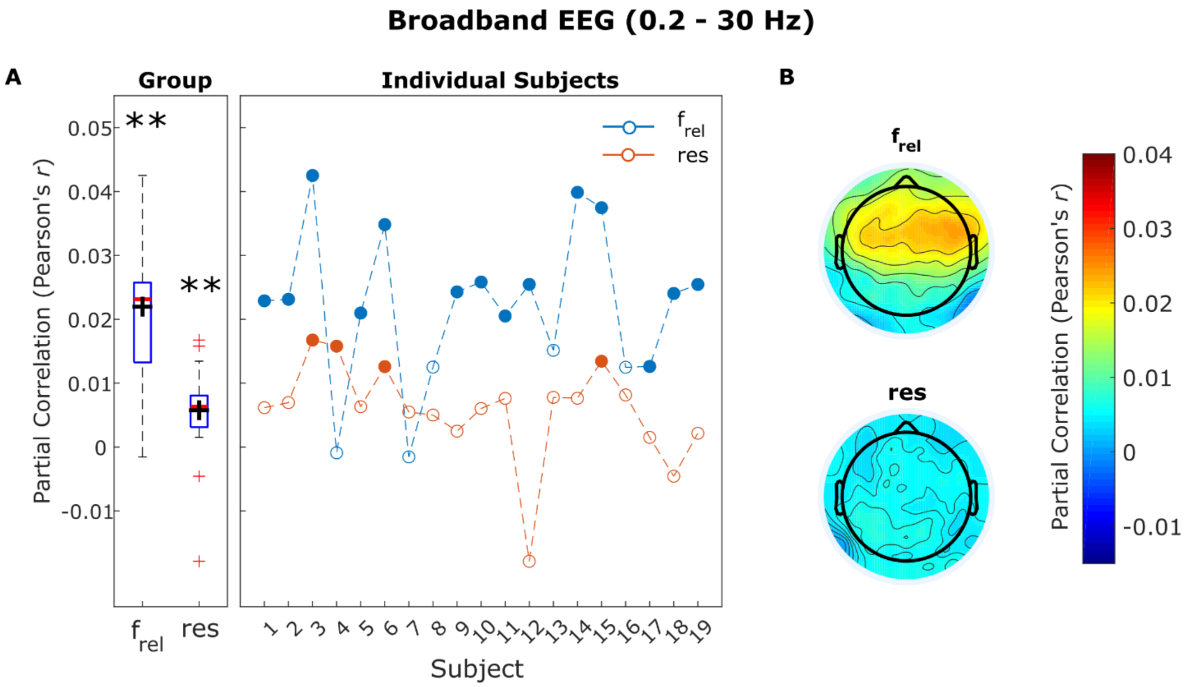


Figure 4-3 | Broadband EEG: partial correlations

*Partial correlations assessing the unique predictive power of two pitch-related measures, f_{rel} and res, after accounting for all acoustic and phonetic measures (average over 30 best channels). (A) Group- and individual-level results; the black crosses in the group plot indicate the mean across all subjects for each feature and asterisks indicate statistical significance (one-tailed permutation test, *p<0.05, **p<0.01). For individual subjects, filled circles indicate statistical significance above chance (one-tailed permutation test, p<0.05). (B) Topographic distribution of the partial correlations for pitch-related measures across all subjects*

4.3.2 Relative pitch tracking is specific to delta phase, but all other features are indexed in both delta and theta phase

To more precisely identify which components of the EEG signal best encode pitch-related features, we computed the power and phase of delta, theta, alpha, and beta band EEG. We repeated the aforementioned regression analysis, mapping our stimulus features to each of these EEG components. We found that, consistent with speech processing literature, our measures were better reflected in the delta and theta phase components than in all other analytic components, with f_{rel} primarily reflected in delta phase. We therefore focused on the delta and theta phase components.

In the delta band, all features could individually predict EEG phase better than chance on a group level (one-tailed permutation test, *env*: p=9.99e-4, *s*: p=9.99e-4, *f*: p=9.99e-4, f_{rel}: p=9.99e-4, and *res*: p=9.99e-4). As before, we computed the partial correlations of the two pitch-related measures. Both features had significant partial correlations (f_{rel}: p=9.99e-4, and *res*: p=9.99e-4), suggesting that they contributed uniquely to predicting delta phase beyond the other features. On an individual subject level, we found significant partial correlation values in 16 of the 19 subjects for f_{rel} and 5 of the 19 subjects for *res*. The group and individual partial correlations of the pitch-related measures are shown in Figure 4-4A.

In the theta band, again all features could individually predict EEG phase better than chance on a group level (one-tailed permutation test, *env*: p=9.99e-4, *s*: p=9.99e-4, *f*: p=9.99e-4, f_{rel}: p=9.99e-4, and *res*: 9.99e-4). However, when we partialled out the contribution of all other features from the EEG predicted by the pitch-related measures, we found that of the two measures, only *res* had a unique contribution to predicting theta phase (p=9.99e-4) – the partial correlation of f_{rel} was not significantly greater than chance (p=0.076). On an individual subject level, we found significant partial correlation values in 3 of the 19 subjects for f_{rel} and 13 of the 19 subjects for *res*. Partial correlations for theta phase are shown in Figure 4-4B.

We examined the topographic distributions of the partial correlations on the scalp (that is, how well data on different channels could be predicted) for delta and theta phase (Figure 4-4C and D). Based on the topographic plots, f_{rel} appears to be more strongly encoded in mid-frontal channels in delta phase EEG, but there were no visible clusters with higher partial correlations for f_{rel} in the theta band. A different pattern of results was observed for *res* – there were no visible clusters with higher partial correlations in delta phase EEG, but two clusters (one on each hemisphere) were noticeable in the theta band.

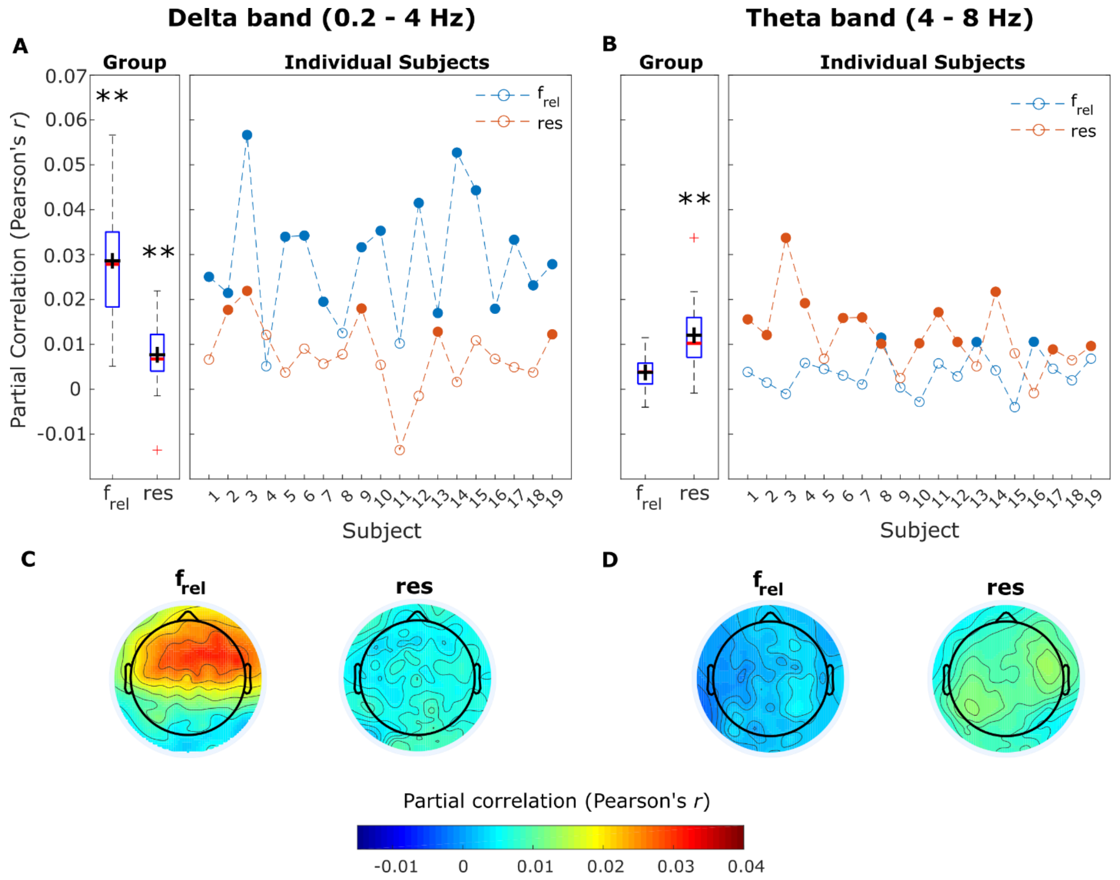


Figure 4-4 | Delta and theta band: partial correlations

*Group and single-subject level partial correlations (Pearson's r; average over 30 best channels) of pitch-related features when predicting the phase for the (A) delta and (B) theta frequency bands of the EEG signal, after controlling for all other measures. Asterisks in the group plots indicate statistical significance (one-tailed permutation test, *p<0.05, **p<0.01) and black crosses indicate the mean across all subjects. For individual subjects, filled circles indicate statistical significance above chance (one-tailed permutation test, p<0.05). Also shown are topographic distributions (C & D) of the partial correlations to delta and theta phase, respectively.*

As well as examining how well data on each channel can be predicted by TRF models, one can also visualize the TRF model weights themselves with a view to understanding what channels show a relationship to the various speech features and at what time-lags those relationships occur (Crosse et al., 2016). This is because the TRF reflects how the instantaneous value of a stimulus feature impacts upon the response. We examined the TRFs of the pitch-related features, focusing again on the phase of delta and theta EEG (Figure 4-5). To control for the effect of all other features, we computed a joint TRF model using all acoustic/phonetic and pitch features and extracted only weights corresponding to our pitch-related measures from the model.

For delta phase, the TRF of f_{rel} was a monophasic signal which peaked in power at ~160ms. The TRF of *res* had a smaller magnitude and peaked later (~200ms) but for comparison, we show the topographic distribution of TRF weights for both measures at ~160ms in Figure 4-5A. The two measures display distinct topographies, with fronto-central positivity for f_{rel} and bilateral fronto-temporal positivity and posterior negativity for *res*. For theta phase, the weights for f_{rel} were small in magnitude; since this measure does not significantly predict theta phase, we did not plot the topography of its weights. The TRF for *res* was a quadriphasic signal that peaked at ~30ms, ~110ms, ~190ms and ~260ms. At these times, the weights were highest in magnitude at bilateral fronto-temporal regions.

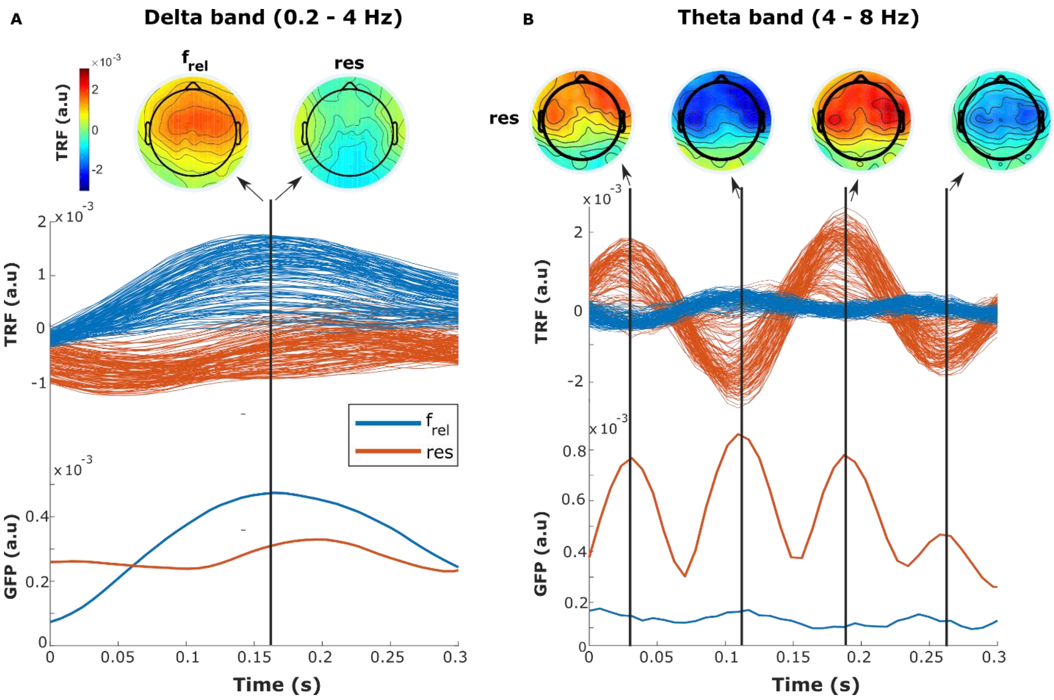


Figure 4-5 | Delta and theta band: TRF weights

TRF weights of all channels (averaged over subjects) that reflect the mapping of individual pitch-related features (obtained by fitting a joint model to all features then extracting weights corresponding to pitch-related measures) to (A) delta phase and (B) theta phase EEG. Topographic distributions of the weights are shown at ~0.16s for f_{rel} *and res in delta phase, and at ~0.03s, ~0.11s, ~0.19s and ~0.26s for res in theta phase (not shown for* f_{rel} *since it does not contribute uniquely above chance in theta). These time points correspond to the peaks in the global field power (GFP) – shown below the TRFs.*

4.3.3 Noise-vocoded speech: Validation that f_{rel} indexes pitch-specific activity

As a check that our pitch-related measures reflect pitch-specific activity, we regressed acoustic and pitch-related measures extracted from clean speech to EEG data (delta and theta phase) recorded during experiment 2. As mentioned above, this experiment involved participants listening to both clean and noise-vocoded versions of the same speech segments. Partial correlations coefficients are shown in Figure 4-6.

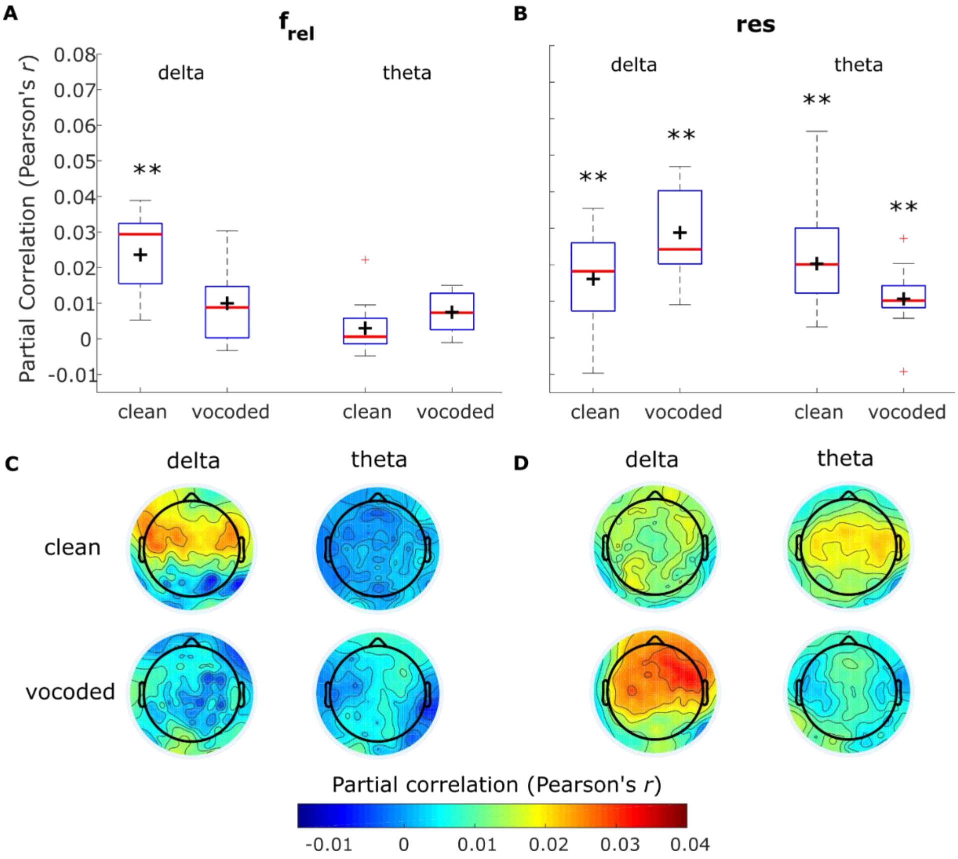


Figure 4-6 | Clean and vocoded speech: partial correlations

*Partial correlations showing how well our pitch-related measures extracted from clean speech predict EEG that was recorded as subjects listened to clean and vocoded-versions of speech snippets. Group-level distributions (average across best 30 channels; but using all channels produced the same pattern of results) are shown for delta and theta phase for (A) f_{rel} and (B) res. Black crosses indicate mean across all subjects, and asterisks indicate significance (one-tailed permutation test; *p<0.05, **p<0.01). The topographic distributions of these partial correlations are also shown in (C) and (D).*

Consistent with our above results, we found significant tracking of f_{rel} in delta band for clean speech (one-tailed permutation test; p = 9.99e-4), but no tracking of this measure in vocoded speech (p=0.1918). There was also no significant tracking of f_{rel} in either clean or vocoded speech in theta band (p=0.9411 and p=0.1918, respectively). In contrast, *res* was present in clean and vocoded speech in both frequency bands (delta: p=9.99e-4 and p=9.99e-4, theta: p=9.99e-4 and p=0.0020, respectively). In delta band, vocoded speech was tracked significantly better than clean speech (two-tailed Wilcoxon signed rank, z=2.3412, p=0.0192); the reverse trend was observed in theta band although this was not significant (z=-1.852, p=0.0640).

4.3.4 No evidence for hemispheric lateralisation of our pitch-related measures

While there is consensus that speech processing occurs bilaterally, the differential roles of the two hemispheres remains controversial. Several auditory processing models (Poeppel, 2003; Zatorre, Belin, & Penhune, 2002) posit (or are compatible with the view) that the right hemisphere is relatively specialised for spectral information in contrast with the left hemisphere’s specialisation for rapidly fluctuating temporal cues. Certain aspects of low-level pitch processing have indeed been found to be lateralised to the right hemisphere (Griffiths et al., 1999; Johnsrude, Penhune, & Zatorre, 2000; Zatorre, Evans, & Meyer, 1994), but lesion and neuroimaging studies have suggested a less clear-cut picture with respect to prosody-related pitch (refer to Section 4.4). With that in mind, we tested for hemispheric lateralisation in the processing of our pitch-related measures. To do so, we averaged the partial correlations of the delta phase EEG predicted by our pitch-related features with the actual delta phase EEG (that is, corresponding to the results in Figure 4-4) over right and left hemispheres (39 on each side; midline and vertex electrodes were disregarded) for each subject and compared their distributions (Figure 4-7). No significant differences were observed between hemispheres for either pitch-related measure (two-tailed Wilcoxon signed-rank, f_{rel}: z=1.0865, p=0.1365; *res*: z=0.9658, p=0.3341).

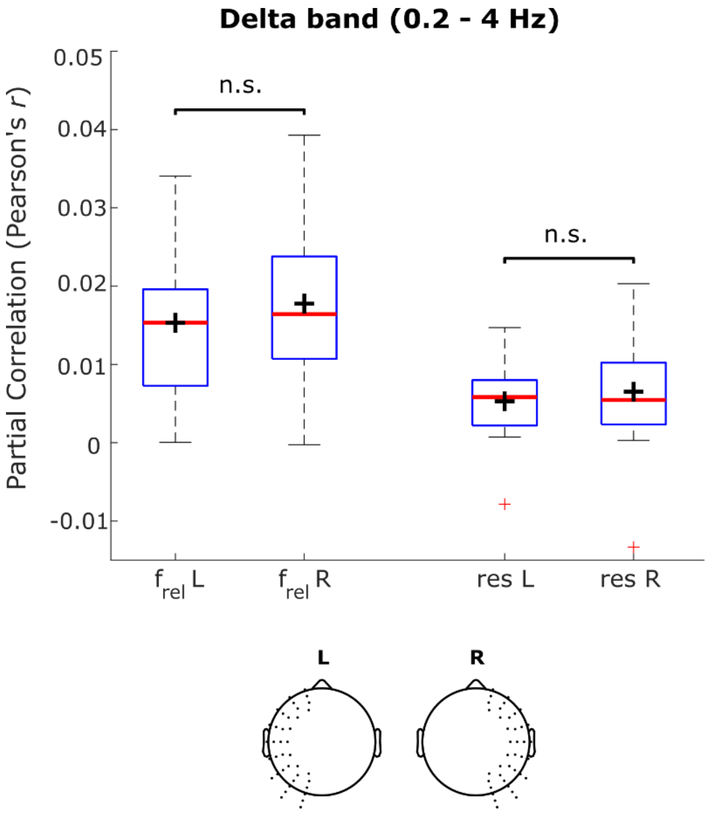


Figure 4-7 | Testing for hemispheric lateralisation

We tested for hemispheric dominance in the processing of pitch features by comparing the average of their partial correlations in delta band across left- and right-sided channels. No significant differences were observed between hemispheres for either of the features (two-tailed Wilcoxon signed rank test (R vs L); f_{rel}: z= 1.4890, p= 0.1365; res: z= 0.9658, p=0.3341).

4.4 Discussion

Recent fMRI and ECoG studies have shown that there are cortical pitch regions which respond primarily to resolved harmonics (Norman-Haignere et al., 2013), and that speaker-normalised relative pitch is encoded in high gamma activity in cortex (Tang et al., 2017). In this chapter, we looked to isolate indices corresponding to these pitch-processing phenomena in scalp-recorded low-frequency EEG during listening to continuous narrative speech. We computed and regressed two pitch-related features – harmonic resolvability (*res*) and relative pitch (f_{rel}) – to ongoing EEG activity (broadband and analytic components of various frequency bands). To determine if they contain unique predictive power, we accounted for three previously established acoustic and phonetic representations: the temporal envelope, a commonly-used measure of the stimulus that captures the coherent fluctuations in time of multiple acoustic features, the spectrogram, which has been shown to model acoustic energy beyond envelope, and phonetic features, which captures some of the acoustics as well as higher-level categorical information.

We found evidence for the encoding of our resolvability feature in broadband EEG and in its delta- and theta-phase analytic components. Combining this measure with our other acoustic/phonetic features to predict broadband EEG led to a significant improvement over using only the joint acoustic/phonetic model. Moreover, partial correlation analysis found that this feature contributed unique predictive power on a group level in broadband EEG, as well as in delta and theta phase. It was particularly prominent in theta phase, where it was significantly greater than chance for the majority of subjects. The question then arose as to whether the unique contribution of our resolvability measure reflects *pitch-specific* activity. This was a concern particularly because we found the time course of the resolvability measure to be highly correlated with the temporal envelope. Indeed, it is evident from the large disparity in the prediction accuracies of *res* before and after controlling for other measures that the tracking of this measure overlaps markedly with these other features. And, given that the acoustic features we included are almost certainly imperfect representations of all coherent low-level fluctuations in the stimulus, it could be that the unique predictive power of *res* is simply due to it capturing some of these residual shared fluctuations. Furthermore, we found this measure to be tracked in both EEG recorded as subjects listened to clean speech as well as its noise-vocoded counterpart (which has degraded pitch). In fact, it was significantly more predictive of EEG for noise-vocoded speech than clean speech in delta phase. This adds to the complexity of pinpointing precisely what the tracking of this measure means.

We were initially motivated to test the resolvability measure based on compelling fMRI findings (Norman-Haignere et al., 2013). Even so, our inability to clearly disambiguate this measure is in no way contradictory to their results. For one, they were able to isolate small, pitch-sensitive regions of auditory cortex on the basis of fMRI's excellent spatial resolution. EEG is lacking in that respect, but we were hoping to leverage EEG's superior temporal resolution to the same effect. Unfortunately, this proved to be complicated simply because many features of speech –resolvability being one of them – covary in time. It may well be that the unique contribution of resolvability to EEG prediction that we have shown contains some pitch-specific activity, but it appears to be too confounded to be a useful measure for indexing pitch processing per se. Nonetheless, it could still be potentially useful in applications where it is imperative to extract as much stimulus-related signal as possible from EEG.

With regards to our relative pitch measure, combining it with the joint acoustic/phonetic model led to a significant improvement in prediction accuracy in broadband EEG. It also portrayed a topographic distribution of prediction accuracies that was markedly different from that of the other measures. And, unlike acoustic/phonetic features and resolvability, the unique contribution of relative pitch was found to be bandlimited to delta phase. The suprasegmental information carried by relative pitch in speech typically fluctuates over intervals of several hundred milliseconds to seconds, so delta phase tracking is consistent with parsing and integrating information at this rate. To validate that this measure is reflective of pitch-specific activity, we compared the tracking of this measure in EEG when subjects listened to clean and noise-vocoded speech. As in Experiment 1, we found significant tracking of relative pitch in delta band for clean speech; but, crucially, there was no tracking for vocoded speech. Together, these results give us confidence that this feature indexes pitch-specific processing.

Our finding complements that of Tang *et al*. (2017), showing that relative pitch is also reflected in low-frequency EEG phase. This is non-trivial as studies examining both phase-locked low-frequency (event-related potentials, ERPs, which are somewhat analogous to the TRF) and non-phase-locked high gamma EEG/ECoG activity (Crone, Boatman, Gordon, & Hao, 2001; Crone, Sinai, & Korzeniewska, 2006; Edwards et al., 2009; Engell & McCarthy, 2011) have found different spatial and temporal profiles for the two measures, suggesting that they result from distinct physiological mechanisms. Although the unique contribution of f_{rel} to predicting EEG appears to be small relative to overall acoustic/phonetic tracking, this was not unexpected as Tang *et al*. (2017) also found only a small percentage of electrodes within auditory cortex to be tuned to relative pitch in comparison with those tuned to the phonetic content of sentences.

A lot of EEG/MEG speech research in recent years has focused on the neural tracking of the temporal envelope and its functional roles (Ding & Simon, 2014). The temporal envelope of speech represents the slow amplitude fluctuations in the acoustic signal and conveys important markers for segmenting linguistic features (syllables in particular) (Rosen, 1992). Indeed, behavioral studies have shown that the envelope contains information essential for speech intelligibility (Drullman et al., 1994a; Drullman et al., 1994b). Given the myriad cues represented in the temporal envelope, the precise function(s) of the observed envelope tracking response – whether it simply reflects passive general auditory encoding or specific language-related processing as well – remains debated. There has nonetheless been some empirical evidence that the measure correlates with speech comprehension (Ahissar et al., 2001; Molinaro & Lizarazu, 2018; Peelle, Gross, & Davis, 2013), suggesting that it relates to both acoustic-phonetic and linguistic processing. Our results here, showing a pitch-specific measure that is dissociable from the envelope, as well as the spectrogram and phonetic features, posits a parallel neural channel for processing meaning conveyed by prosodic pitch.

General auditory processing models (Poeppel, 2003; Zatorre, Belin, & Penhune, 2002) predict a right hemispheric bias for pitch processing. But neural correlates of prosodic pitch have only been observed to be right lateralised within the context of certain tasks, and typically when lexical content is no longer identifiable (Baum & Pell, 1999; Gandour et al., 2004; Kotz et al., 2003; Kreitewolf et al., 2014; Meyer et al., 2002; Meyer et al., 2003; Peretz, 1990; Plante, Creusere, & Sabin, 2002; Sammler et al., 2015; Witteman et al., 2011; Zatorre et al., 1992). Left lateralization has also been observed within the context of contrasting tonal and non-tonal language processing (Gandour et al., 2004). Zatorre and Gandour (2008) surmised that multiple local asymmetries underlie prosodic pitch processing, with right-hemisphere activity reflecting lower-level, domain independent pitch processing (in accordance with general models) and left-hemispheric activity reflecting higher-level language-related processing. The emergence of lateralization effects is therefore dependent on the interaction of these processes in a specific listening situation. Here, in a scenario where subjects listened to continuous natural (non-tonal) speech, much like they would do in real life, we did not find any hemispheric bias for relative pitch (nor for resolvability, for that matter). That is, there was no significant difference in how accurately relative pitch could predict delta phase EEG activity from the right and left hemispheres. EEG's low spatial resolution precludes precise anatomical localisation; thus, our result should be interpreted with caution as it could be the case that the effect is simply too small to be detected with EEG. Nonetheless, it lends support to the notion that prosodic pitch processing in the context of natural speech is subserved by a complex network spanning both hemispheres (Zatorre and Gandour, 2008).

Like ECoG, EEG has excellent millisecond temporal resolution that lends itself to the task of understanding how rapid speech dynamics are encoded, but with the added advantages of being portable and non-invasive. These properties facilitate the application of any findings to cognitively-controlled smart hearing devices (Lunner, Rudner, & Rönnberg, 2009) and speech processing studies in specific cohorts. For instance, individuals with autism often display disordered or unusual prosody (McCann & Peppé, 2003). Neural processing of prosody in individuals with autism remains an under-researched topic and an isolated index of relative pitch processing in continuous natural speech as shown here could help shed light on the nature of their prosodic deficits.

4.5 Summary

In this chapter, we investigated if we could isolate low-frequency EEG indices of pitch processing of continuous narrative speech from those reflecting the tracking of other acoustic and phonetic features. Harmonic resolvability was found to contain unique predictive power in delta and theta phase, but it was highly correlated with the envelope and tracked even when stimuli were pitch-impoverished. As such, we are circumspect about whether its contribution is truly pitch-specific. Crucially however, we found a unique contribution of relative pitch to EEG delta phase prediction, and this tracking was absent when subjects listened to pitch-impoverished stimuli. This finding suggests the possibility of a processing stream for prosody that might operate in parallel to acoustic-linguistic processing. Furthermore, it provides a novel neural index that could be useful for testing prosodic encoding in populations with speech processing deficits and for improving cognitively-controlled hearing aids.

Chapter 5 Investigating the effects of selective auditory attention on pre-lexical and prosodic pitch processing

5.1 Introduction

As mentioned in earlier chapters, cortex is posited to extract meaning from speech via a hierarchy of increasingly abstract intermediary representations, with feedforward and feedback connections between the levels. But this is complicated by the fact that everyday acoustic scenes often contain concurrent sources besides the speech stream of interest to the listener. As a consequence of the brain's limited processing capacity, not all sources present can be exhaustively analysed. In Section 2.4, we reviewed the literature on speech comprehension in multi-talker scenarios (i.e. the cocktail party problem), where people selectively attend to one source. Much debate has revolved around the issue of whether attention to speech operates at an early or late stage of processing. Behavioural studies have made important contributions to this debate: it was found that in more naturalistic experiments using continuous speech, subjects typically perform no better than chance when tasked with answering questions on the ignored stream (e.g. Mirkovic et al., 2016; Mirkovic et al., 2015; Power et al., 2012); but it has also been observed that subjects are aware of certain aspects of the unattended stream (Cherry, 1953; Moray, 1959). Early behavioural findings led to several different models of attention (Broadbent, 1958; Deutsch & Deutsch, 1963; Treisman, 1964). Nonetheless, behavioural studies are limited in their ability to probe how ignored speech is represented in cortex.

Recently, the development of encoding/decoding approaches for high temporal fidelity neuroimaging modalities has provided an opportunity to investigate the processing of the acoustic scene through the hierarchy. These techniques have led to the discovery of neural indices of speech representations at various levels – from acoustic to semantic – during listening to continuous natural speech (Ahissar et al., 2001; Brodbeck, Hong, & Simon, 2018; Broderick et al, 2018; Daube, Ince, & Gross, 2019; Di Liberto et al., 2018; Di Liberto, O'Sullivan, & Lalor, 2015; Mesgarani et al., 2014; Tang, Hamilton, & Chang, 2017). Subsequently, a few studies have investigated how some of these measures are modulated by attention through the implementation of a 'cocktail party' paradigm. Here, it has been shown that an estimated TRF to the acoustic envelope of speech is attenuated in the 200-220ms range (Power et al., 2012). More recently, Brodbeck et al. (2018) and Broderick et al. (2018) showed that indices at the lexical and semantic levels are categorically encoded in neural activity only when a speech stream is attended to. These findings also fit with ECoG work showing that only the attended speaker is represented in STG, which is thought to implement a crucial stage in processing at the transition from acoustic to linguistic processing (Mesgarani & Chang, 2012). Together, these results present evidence to support the notion that higher-order regions exhibit greater attentional selectivity.

In spite of these discoveries, the effect of attention at the pre-lexical level remains incompletely explored. This is in part because how speech is represented at the pre-lexical level remains contentious (Section 2.3.4). Namely, some models posit a mapping to phonetic features as an intermediate stage between low-level acoustics and words (McClelland & Elman, 1986; Liberman et al., 1967). However, it remains controversial whether this phonetic/phonemic-level is necessary at all (Lotto & Holt, 2000). Several studies have shown that a phonetic feature representation of speech can predict neural activity (Di Liberto et al., 2015; Khalighinejad, et al., 2017; Mesgarani et al., 2014). But a recent study contended that the EEG/MEG findings in particular can be accounted for by including low-level acoustic feature spaces such as the derivative of the spectrogram, and phoneme onsets (Daube et al., 2019).

In this chapter, we explore how attention affects the phonetic feature representation in comparison to the other acoustic/feature onset representations suggested to account for the same predictive power in Daube et al. (2019) within the context of a two-talker cocktail party paradigm. In accordance with previous findings, if the phonetic feature representation can uniquely predict neural activity and is more strongly moderated by attention than these feature spaces, it will lend support to the notion of it being a distinct and higher-level stage in the hierarchy. As a follow up to our study in Chapter 4 (Teoh, Cappelloni & Lalor, 2019), we also examine how attention affects relative pitch, a continuous percept that underlies intonational prosody, and whether this measure can be leveraged for decoding attentional selection.

5.2 Methods

5.2.1 Subjects

Nine female and five male subjects between the ages of 19 and 30 took part. All subjects were right-handed and spoke English as their primary language. Subjects reported no history of hearing impairment or neurological disorder. Each subject provided written informed consent prior to testing and received monetary reimbursement. The study was approved by the Research Subjects Review Board at the University of Rochester. Some of the data (the first 10 trials) used here were also used as a control condition for the study in Chapter 3, which has been published (Teoh & Lalor, 2019).

5.2.2 Stimuli and Procedures

Subjects undertook 40 one-minute trials in two separate blocks. Stimuli consisted of two Sherlock Holmes stories narrated by a female and a male talker. Silent gaps in the audio exceeding 0.3 s were truncated to 0.3 s in duration.

The streams were filtered using Head-Related Transfer Functions (HRTFs), which characterise how the ear receives a sound from a point in space, taking into account the diffraction and reflection properties of the listener's head and ear. We selected HRTFs that gave rise to the perception that the talkers were at 90 degrees to the left and right of the subject. The HRTFs used were obtained from the CIPIC database (Algazi et al., 2001). Subjects were always instructed to attend to one of the two talkers – a counterbalanced paradigm was employed in which over the course of the experiment, they would have attended to both male and female talkers and at both locations.

Before the experiment, subjects were asked to minimize motor activities and to maintain visual fixation on a crosshair centred on the screen during trials. After each trial, subjects were required to answer four multiple-choice comprehension questions on each of the stories (attended and unattended).

All stimuli were sampled with a frequency of 44 100 kHz and were presented using Sennheiser HD650 headphones and Presentation software from Neurobehavioral Systems (http://www.neurobs.com).

5.2.3 Data Acquisition and Pre-processing

The experiment was conducted in a soundproof room. A Biosemi ActiveTwo system was used to record EEG data from 128 electrode positions on the scalp as well as two electrodes over the mastoid processes (all digitized at 512Hz).

EEG data were re-referenced to the mastoids. Automatic bad channel rejection and interpolation was performed. A particular channel was deemed as bad if the standard deviation of the channel was lower than a third of or exceeded three times the mean of the standard deviation of all channels. In place of the bad channel, data were interpolated from the four nearest neighbouring electrodes using spherical spline interpolation (Delorme & Makeig, 2004). To decrease subsequent processing time, data were down-sampled to 128Hz. Data were filtered between 0.2 – 8 Hz using a Chebyshev II filter. Low delta-band frequencies were included as the study in Chapter 4 found them to be important for the encoding of relative pitch.

5.2.4 Speech Representations

We computed phonetic, acoustic, and pitch-related speech representations of both the attended and unattended streams of speech.

Phonetic:

- Phonetic features (*f*): Phoneme-level segmentation of the stimuli was first computed using FAVE-Extract (Rosenfelder et al, 2014) and the Montreal Forced Aligner (McAuliffe et al, 2017) via the DARLA web interface (Reddy & Stanford, 2015). The phoneme representation was then mapped into a space of 19 features (based on the University of Iowa's phonetics project). The features described the articulatory and acoustic properties of the phonetic content of speech. Based on the start time of each feature, a multivariate time-series binary matrix (19 features by samples) was produced to mark feature onsets
- Feature onsets (*fo*): Univariate vector of the onsets of all phonetic features (i.e. a summary measure of all onsets in *f* without categorical discrimination)

Acoustic:

- Spectrogram (*s*): The GBFB toolbox (Schädler & Kollmeier, 2015) was used to extract the spectral decomposition of the time-varying stimulus energy in 31 mel-spaced bands with logarithmic compressive nonlinearity. The spectrogram representation was computed in this way as this was shown in Daube et al. (2019) to better predict neural activity than other techniques.

- Spectrogram Derivative (*sD*): The temporal derivative of each spectrogram channel was computed and then half-wave rectified. This represents the rate of change of power in each channel.

Prosodic:

- Relative Pitch (f_{rel}): Praat software (Boersma & Weenink, n.d.) was used to extract a continuous measure of pitch (fundamental frequency/absolute pitch) at a sampling rate of 128 Hz. It performs pitch estimation using an autocorrelation method (Boersma, 1993). The z-score of this measure was then computed to obtain speaker-normalised relative pitch.

5.2.5 TRF implementation

The acoustic, phonetic, and relative pitch representations for both attended and unattended speech described above were normalised, grouped into various joint feature sets (described in section 5.3.2) and mapped to the concurrently-recorded 128-channel EEG signals. As in the case of the study in Chapter 4, encoding/forward modelling was performed in which multivariate regularized linear regression (TRF approach) was employed. That is, each EEG channel is estimated to be a linear transformation of the speech features over a range of time lags where the transformation is described by the TRF.

Again, the regularisation parameter (Section 2.6) was tuned using leave-one-out cross-validation. That is, we trained on *n*-1 trials for a wide range of λ values (e.g. 1, $1\text{x}10^2$, … $1\text{x}10^9$), computed the average TRF across trials for each λ, then tested the TRFs on the *n*th trial. This was repeated *n* times, rotating the trial to be tested each time. The λ value that maximized the Pearson's correlation coefficient between the actual and predicted neural response over all trials was selected.

The transformation was computed over time lag intervals of 0 to 300ms, selected based on previous speech-related studies (Di Liberto, O'Sullivan, & Lalor, 2015). We quantified how well each speech representation related to the neural data using leave-one-trial-out cross-validation (as described above), with Pearson's correlation coefficient as our metric of prediction accuracy.

To evaluate whether a feature contributed independently of all other features in predicting the neural data, we also computed the partial correlation coefficients (Pearson's *r*) between the EEG predicted by each measure's model with the actual recorded EEG after controlling for the effects of all other features.

5.2.6 Statistical Testing

To test that a partial correlation coefficient is above chance level, we performed non-parametric permutation testing. The predicted EEG activity for each model's representation was permuted across trials such that they were matched to the actual EEG of a different trial, and partial correlation coefficients were computed, controlling for the effects of all other features. This was done 100 times for each subject to establish a distribution of chance-level prediction accuracies. To perform group-level statistical testing, we generated a null distribution of group means: one prediction accuracy from each subject's individual distribution was selected at random to go into each group mean. This process was repeated 100 times, sampling with replacement for each subject. For comparison between groups (e.g. attended vs unattended for a particular representation), two-tailed Wilcoxon signed-rank testing was used.

5.3 Results

5.3.1 Behavioural Results

Subjects were found to be compliant in carrying out the behavioural task. The average questionnaire accuracy was 73.2 ± 3.2% when subjects were tested on the attended story and 27.2 ± 1.6% for unattended stimuli (theoretical chance level is 25% as there are four possible answers to each question).

5.3.2 Responses to attended talker reflect phonetic and prosodic processing

We assessed the performance of attended and unattended models that were trained to predict EEG responses using different combinations of acoustic and phonetic features spaces extracted from the speech stimulus. The (Pearson's) correlation between predicted and actual EEG time courses, averaged across the same 12 fronto-temporal channels used in Di Liberto et al. (2015) and Daube et al. (2019), was used as our metric of prediction accuracy for each individual model. We performed statistical testing between the results of different models, with the goal of evaluating whether the phonetic features representation contributed predictive power above and beyond acoustic feature spaces –

these results are shown in Table 5-1. We found that including phonetic features significantly improved prediction over the two acoustic feature spaces alone (i.e. *s* & *s+sD*) when subjects were attending to the stimuli; but this improvement was not observed for the unattended stimuli. Additionally, it also improved prediction over phoneme onsets in the attended case, showing that the information on specific feature categories adds predictive power. The spectrogram derivative representation displayed a different pattern of results – including this measure in the feature space with spectrogram (i.e. *s+sD*) improved prediction accuracy over spectrogram alone for both attended and unattended stimuli, suggesting that it is less modulated by attention.

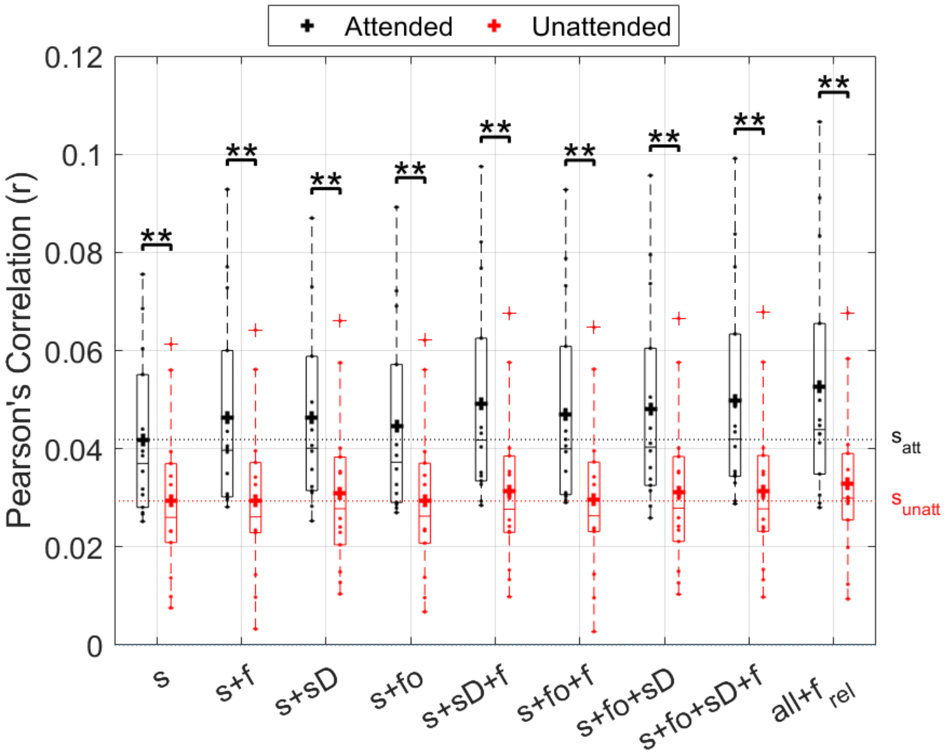


Figure 5-1 | Prediction accuracies for acoustic, phonetic and prosodic measures under varying attentional conditions

*Prediction accuracies for attended and unattended models based on different representational features spaces, as shown on the horizontal axis. For all feature spaces, prediction accuracy was significantly better when stimuli are attended to (**p<0.01, two-tailed Wilcoxon sign-ranked test). The dashed horizontal black and red lines indicate mean prediction accuracies for the attended and unattended spectrogram models respectively.*

	Attended	Unattended
baseline: s		
$s + f$ > baseline	**p = 0.0015, z = 3.1702**	p = 0.3003, z = 1.0358
$s + sD$ > baseline	**p = 0.0023, z = 3.0447**	**p = 0.0019, z = 3.1074**
$s + fo$ > baseline	**p = 9.82e-04, z = 3.2958**	p = 0.1578, z = 1.4125
baseline: $s + sD$		
$s + sD+ f$ > baseline	**p = 0.0043, z = 2.8563**	p = 0.1981, z = 1.2869
baseline: $s + fo$		
$s + fo + f$ > baseline	**p = 0.0023, z = 3.0447**	p = 0.6378, z = 0.4708
baseline: $s + fo + sD$		
$s + fo + sD + f$> baseline	**p = 0.0076, z = 2.6680**	p = 0.4703, z = 0.7219
baseline: $s + fo + sD + f$		
$s + fo + sD + f + f_{rel}$ > baseline	**p = 0.0157, z = 2.4169**	p = 0.1240, z = 1.5380

Table 5-1 | Joint model comparison results

Wilcoxon two-sided signed rank test statistics for attended and unattended speech streams (columns). Each set of rows test a different statistical question, as specified by the baseline. Terms being evaluated are to the left of the '>' symbol. Bolded values

5.3.3 Isolating the unique predictive power of each representation

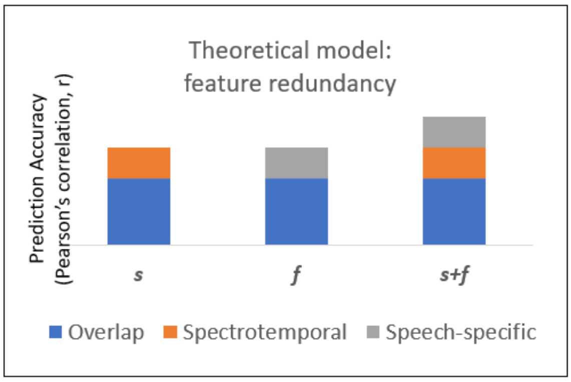


Figure 5-2 | Illustration of feature redundancy

Theoretically, there will be overlapping contributions to the combined model prediction accuracy from the spectrogram and phonetic feature models due to shared acoustic invariance between the instances of a particular phonetic feature. Other feature spaces also share overlapping contributions.

It is important to note that there is redundancy between the various speech representations examined here. Phonetic features (*f*) has some overlap with the spectrogram (*s*) in that if every phoneme is always spoken the same way, then the two representations would be equivalent (Figure 5-2). The phoneme onset (*fo*) representation indicates the same onset times as the phonetic feature representation (*f*), except that it does not contain information on the different feature categories. The spectrogram derivative (*sD*), a measure of the positive temporal rate of change of power in each channel of the spectrogram, would have some peaks that overlap with those of the phoneme onset (*fo*) representation (e.g. after every silent period in the stimulus). Given these redundancies, we were interested in more clearly isolating the *unique* contribution of each feature. To do so, we trained models using each individual representation, then employed a partial correlation approach to control for the predictions of all other representations. The unique predictive power of each model is shown in Figure 5-3A (shown here for an average across 12 channels, although including all channels revealed the same pattern of results). All feature models – attended and unattended – contributed uniquely at above chance level on a group level (p=0.0099 for all models; non-parametric permutation test). Nonetheless, on an individual level, only a minority of subjects (5 out of 14) were found to predict EEG at above chance for the *f* and f_{rel} unattended models (subjects for which significance is achieved are depicted with circular markers in Figure 5-3A). For all other representation models (attended and unattended), a majority of subjects (i.e. at least 7 out of 14) could predict EEG at above chance. Additionally, we found a significant effect of attention for phonetic features and relative pitch, but not for any of the other representations. The topographic distributions for the unique predictive power of each model are shown in Figure 5-3A. Consistent with findings from Chapter 4, the activity unique to attended f_{rel} appears to be encoded in mid-frontal channels.

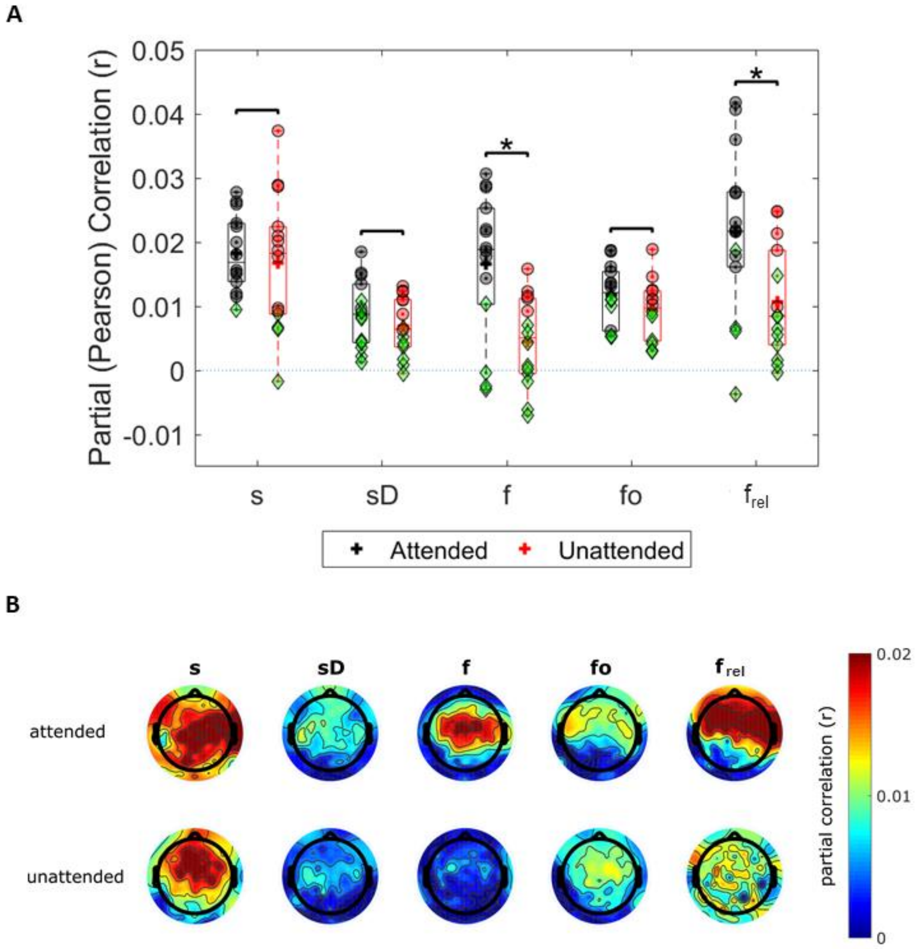


Figure 5-3 | Partial correlations of individual features

*(A) Unique predictive power for the acoustic, phonetic and prosodic feature spaces, under varying attentional conditions. The features are as labelled on the x-axis. All models could significantly predict EEG better than chance on a group level (permutation test; p<0.05). Significance at an individual subject level is indicated by a circular marker (permutation test; p<0.05); non-significance is indicated by a green diamond marker. Statistical testing was also carried out to identify attentional effects (two-tailed Wilcoxon signed rank; *p<0.05) (B) Topographic distribution of partial correlations*

5.3.4 Decoding attentional selection via reconstruction of relative pitch

The decoding implementation of the TRF framework is an effective method for reconstructing continuous measures, as evident from the efficacy of envelope reconstruction for decoding attentional selection. Given that relative pitch is a continuous signal that can be estimated in a relatively straightforward manner from the speech signal (e.g. using an autocorrelation method, as employed in Praat software), and its neural representation is dissociable from the acoustic envelope and shown above to be modulated by attention, we wondered whether we could exploit relative pitch reconstruction to improve decoding performance. To test this, we employed the strategy laid out in Figure 3-2 – that is, we computed attended decoders mapping from EEG to the relative pitch representation for *N*-1 trials, took the average over all decoders, and used it to reconstruct relative pitch for the *N*th (i.e. left out) trial. The correlations (Pearson's) between the reconstructed feature and the actual attended and unattended relative pitch signals were then evaluated. This was done *N* times, rotating the trial to be tested. A linear discriminant analysis (LDA) classifier was used to decode attentional selection based on the attended and unattended relative pitch correlation values (we also tested using an SVM classifier, but the results were not significantly different). As would be expected based on our forward modelling results, decoding accuracy was greater than chance level for a majority of subjects (all except two; binomial test at a 5% confidence level). This is shown in Figure 5-4.

The above procedure was then repeated using the acoustic envelope (computed using the method described in Chapter 3.2.2), as well as a combination of the two representations (that is, we used all four accuracy metrics – the correlations between the reconstructed envelope and attended and ignored envelopes and the correlations between the reconstructed relative pitch and attended and ignored relative pitch – as features for our classifier). These results are also shown in Figure 5-4. Of the two individual decoders, the envelope decoder performed significantly better than relative pitch (two-tailed Wilcoxon signed rank, z=3.1925, p= 0.0014). Combining the two representations resulted in a greater average group decoding accuracy than using envelope reconstruction alone, but this improvement was not significant (two-tailed tailed Wilcoxon signed rank, z=1.9295, p=0.0537). On an individual subject level, mean decoding accuracy for the combined decoder was better than envelope reconstruction for 6 of the 14 subjects, but this improvement was only found to be significant for one subject (Sub 14; one-sided binomial test, p=0.0148). This test is nonetheless conservative and requires large differences between conditions to achieve significance given the small sample size (n=40). The decoding accuracy of one subject (Sub 12) decreased, but this result was not significant.

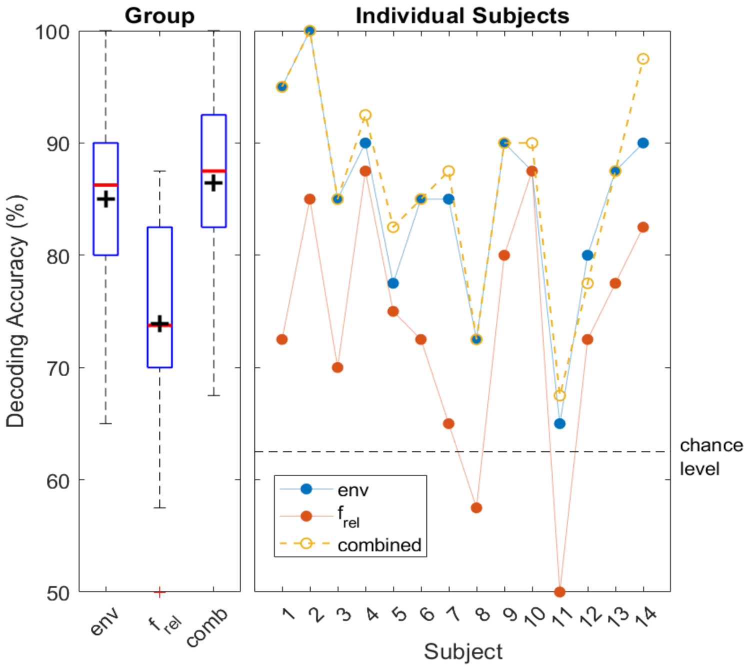


Figure 5-4 | Decoding attentional selection based on the reconstruction accuracies of different features

Decoding accuracies when using (1) envelope reconstruction, env, (2) relative pitch reconstruction, f_{rel}, and (3) a combination of the two features, comb/combined, as inputs to an LDA classifier for decoding the attended speaker. The subplot on the left depicts box plots for the three decoders; group means are indicated by the black crosses. The subplot on the right shows individual subject results.

5.4 Discussion

This study looked at how selective attention modulates pre-lexical and prosodic processing. We recorded EEG from subjects as they listened to trials in which two concurrent talkers (male and female) – simulated to be spatially-separated using HRTFs – narrated stories. In each trial, subjects were instructed to attend to one of the two talkers. We transformed our speech stimuli to phonetic and acoustic feature spaces as described by Di Liberto et al. (2015) and Daube et al. (2019), as well as the relative pitch representation described in Chapter 4. We then constructed joint feature spaces – by starting with the spectrogram as the baseline, and incrementally adding features to this matrix – for the attended and unattended stimuli. These feature matrices were mapped to the EEG to investigate the effect of attention on the various representations. To more clearly isolate the unique contribution of each feature, we also performed partial correlation analysis.

Unlike in Daube et al. (2019), we found that the phonetic feature representation contributed predictive power above and beyond acoustic spaces and phoneme onsets alone when subjects are attending to a talker – this was the case for both the joint modelling approach, and when partial correlation analysis was performed to more clearly isolate the unique contributions of each representation. In the case of the ignored speech stream, the joint modelling method found that adding the phonetic feature representation to acoustic feature spaces did not contribute predictive power. However, we found a weak but significant contribution above chance on a group level, as well as a significant contribution above chance for a minority of subjects on an individual level based on the partial correlation analysis. It is possible that this weak unique predictive power is simply due to some shared invariance in the underlying acoustics between phoneme instances that is not accounted for by the acoustic feature spaces considered here. That is, it could be that our phonetic feature representation is capturing some of this residual shared spectro-temporal information. Nonetheless, with regards to the debate on early versus late selection, our results cannot definitively rule out phonetic-level processing of the unattended speech.

We also found that the *unique* contributions of the feature spaces suggested to explain away the predictive power of phonetic features in Daube et al. (2019) – the spectrogram derivative and feature onsets – were not significantly modulated by attention. This finding – that the phonetic feature representation contains unique predictive power that is more strongly modulated by attention than the aforementioned (theoretically, lower-level) feature spaces – supports the notion that it is a distinct and higher-level stage in the speech processing hierarchy, and is in accordance with previous studies showing that higher-order regions exhibit greater attentional selectivity.

It is possible that we found explanatory power for phonetic features in the attended talker condition because our experiment involved subjects attending to two different talkers (a male and a female), while the experiment in Daube et al. (2019) only involved listening to a single talker. Having different talkers would lead to less overlap between the acoustic and phonetic feature spaces, as there would be larger spectro-temporal differences between instances of a particular phonetic feature category.

Of course, our results cannot definitively prove that the phonetic feature representation is encoded in neural activity as measured by low-frequency EEG during everyday listening. There are other acoustic transformations of the auditory stimulus that have not been considered beyond the spectrogram and its derivative, as well as other theorized intermediate representations apart from phonemes/phonetic features that have not been tested. It is also conceivable that the brain may adjust its speech processing strategy when performing a difficult cocktail party task in comparison to the scenario of listening to a single talker. Indeed, it is known that different kinds of masking sounds place different demands on cognitive resources (Evans, et al., 2016). Future work could be to test if training across many talkers could increase the unique predictive power of the phonetic feature representation in comparison to training on a single talker in the context of a continuous natural speech experiment without distractors.

As an initial check of our phonetic feature representation, we computed individual phonetic feature TRFs for the attended talker. We found phonetic feature TRFs that appeared visually similar to the phoneme related potentials (PRPs) shown in Khalighinejad et al. (2017) (Figure 5-5). That is, they had similar temporal profiles, and also displayed diverse component timings across features. This is not unexpected given that the TRF is analogous to computing an ERP. These findings are also consistent with Di Liberto et al. (2015), where TRFs for the different feature categories (in a single talker/clean speech listening scenario) displayed significant differences and were discriminable at various latencies. Future work could be to explore how attention modulates neural responses to individual features and the discriminability across categories in time.

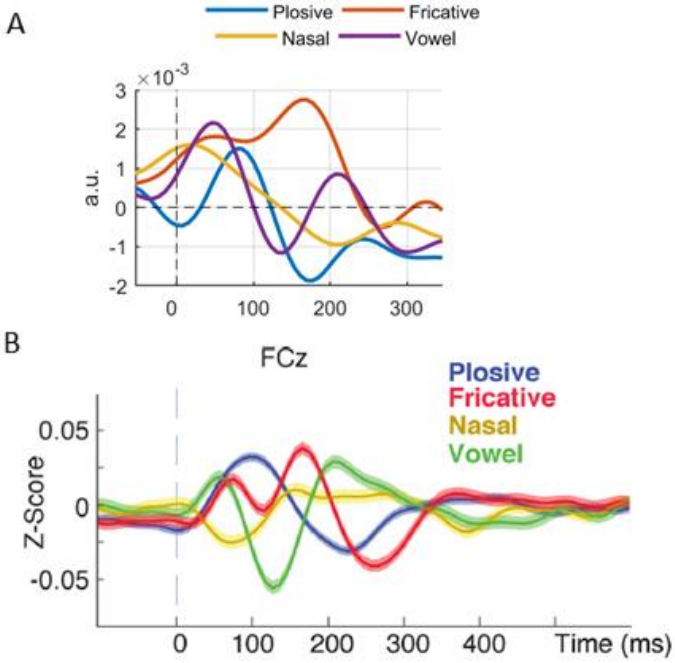


Figure 5-5 | Comparison of phonetic features TRFs and Phoneme Related Potentials

(A*) TRFs and (B) Phoneme-related Potentials (adapted from Figure 2A in Khalighinejad et al., 2017) for four phonetic feature groups at location FCz on the scalp*

In a follow-up to our finding in Chapter 4 that delta band EEG tracks relative pitch, we were interested in investigating how this measure is modulated by attention. We found a significant attentional effect for the encoding of relative pitch. Given this result, we then tested whether this feature could be used for the decoding of attentional selection, and if the incorporation of relative pitch into the decoding algorithm established in O'Sullivan et al. (2015) could improve its performance. On its own, relative pitch could decode attentional selection at above chance, but at a significantly lower accuracy percentage than the acoustic envelope. When combined with the envelope (using an LDA classifier), a small improvement was found for some subjects, although the improvement was only significant for one subject (albeit based on a conservative statistical test). Future work could focus on optimising the algorithm for combining different representational features to determine if we can better leverage their unique predictive powers.

Despite its sensitivity to attentional effects, we did not consider including phonetic features into the decoding algorithm as it has a sparse binary feature matrix which is not ideal for our stimulus reconstruction approach and it is also currently less practical to obtain the phonetic feature representation from the speech signal in real-time – our pipeline here necessitates speech-to-text transcription and forced alignment. Nonetheless, this could be a consideration in the near future when new approaches for phoneme alignment and stimulus-response mapping (methods such as canonical correlation analysis – Section 6.4 – may offer an avenue) are established.

5.5 Summary

In this chapter, we investigated the effects of selective auditory attention on pre-lexical and prosodic processing of speech. The former question was motivated by the longstanding debate as to whether the brain computes an intermediate phoneme representation in going from sound to the lexicon. There have been recent conflicting EEG/MEG findings, with one study demonstrating the encoding of phonetic features, but a subsequent study suggesting that the observed predictive power could be explained by the inclusion of other acoustic features. Given previous evidence that higher order areas in the hierarchy are more affected by attention, we hypothesised that attentional effects might enable us to better disambiguate between acoustic and phonetic (if present) contributions to predicting EEG. We found unique predictive power for the phonetic feature representation above and beyond all other features, and this power was modulated by attention. We also found unique predictive power for features posited to explain away the phonetic feature representation, but they were differently modulated by attention (i.e. they were not significantly affected). Lastly, we showed that prosodic pitch is also modulated by attention, and that this effect manifests itself in the reconstruction accuracy of the measure when a decoding approach was employed. We could leverage this to slightly improve decoding performance over envelope reconstruction alone for several subjects.

Chapter 6 General Discussion

As set out in Chapter 1, speech comprehension is a complex task, involving the interaction of bottom-up and top-down processes. One line of inquiry within this topic concerns precisely how information at each processing stage is encoded in the path from sound to meaning. There is evidence that cortical areas for speech processing are hierarchically organized, with higher levels processing increasingly abstract representations. But understanding the precise computations at each level within this framework remains a work in progress. Another main subarea concerns how selective auditory attention modulates speech processing, and in particular, the extent to which unattended speech undergoes processing. With regards to this, it has been shown that when one of two concurrent talkers is attended to, there are distinct representations of the talkers' acoustic envelopes in the listener's cortex, and that the representation of the unattended talker's envelope is diminished. This seminal finding has given rise to further research along several lines, including studies employing a similar framework to investigate attentional effects along the speech processing hierarchy, and studies developing the idea of exploiting this signature of attention as a means of decoding the user's intended listening direction in 'smart' devices like steerable hearing aids.

The three specific aims of this thesis, set out in Section 1.2, relate to the research subareas outlined above. This chapter will summarize our findings and discuss them with regards to the broader field. Section 6.1 will address decoding attentional selection in a multi-talker scenario using acoustic envelope reconstruction and our efforts to incorporate measures beyond the envelope into the framework. Section 6.2 will discuss our finding that low-frequency EEG tracks relative pitch in cortex in the context of current speech comprehension models. Section 6.3 will address cortical speech processing at the pre-lexical level – evidence for and against a phonetic-level representation in cortex, and how results from our selective attention study contributes to the debate.

6.1 Decoding attentional selection – moving talkers, and looking beyond envelope reconstruction

The acoustic envelope of speech is a low-level measure of acoustic energy that carries important cues for understanding speech. A recent seminal discovery is that cortical activity tracks this measure when people listen to continuous natural speech (Ahissar et al., 2001; Luo and Poeppel, 2007; Aiken and Picton, 2008), and that it can be reconstructed from low-frequency EEG using modelling approaches like the TRF (O'Sullivan et al., 2015). Importantly, this tracking is modulated by attention (Ding & Simon, 2012; Nima Mesgarani & Chang, 2012b), which is also reflected in the accuracy of the acoustic envelope reconstruction (Ding & Simon, 2012; Zion Golumbic et al., 2013).

A framework for exploiting this finding to decode selective auditory attention within the context of a multi-talker scenario was laid out in O'Sullivan et al. (2015). Subjects were asked to attend to one of two talkers in a dichotic listening test, and high-density scalp EEG was concurrently recorded. An attended decoder that mapped from the attended speaker's speech envelope was then trained using the TRF approach. It was subsequently shown that this decoder could be used to reconstruct the amplitude envelope from EEG data of an unseen trial, and that on average, this reconstruction correlates better with the attended talker's actual envelope than that of the ignored talker (a decoding accuracy of ~89% was achieved for 60-second-long trials).

The above discovery provides a potential avenue for the implementation of 'smart' hearing devices – the general premise of which is to enhance the user's ability to recognise speech by steering a directional microphone system towards the talker of their choice. With an eye towards real-world application of the framework in O'Sullivan et al. (2015), several studies have built upon it by looking into practical considerations such as the effects of background noise and how it might be suppressed (Das et al., 2018; O'Sullivan et al., 2017; Van Eyndhoven, Francart, & Bertrand, 2017), reducing the number of channels and training data required for decoding (Mirkovic et al., 2015), optimizing the decoding algorithm using other modelling approaches (Akram, Simon, & Babadi, 2017), and testing the decoder framework in a binaural listening environment (Das, Biesmans, Bertrand, & Francart, 2016).

In this thesis, we contributed towards this goal in several ways. In Chapter 3, we considered the efficacy of the decoding approach established in O'Sullivan et al. (2015) in a scenario where talkers are not stationary. Subjects were tasked with attending to a single talker even as the attended and ignored talkers switched locations. We were motivated to test this as real-world situations are often not stationary, and because previous research on auditory scene analysis has found evidence that even non-task related features of an object can influence performance on a task (see Section 2.4). Our study found that the decoding approach established in O'Sullivan et al. (2015) is robust to switches in talker location – that is, there was no significant decrease in decoding accuracy in comparison to the control stationary speaker condition. This finding complements previous studies, showing the practicality of decoding attention using envelope reconstruction in a more complex, non-stationary environment. Indeed, our result taken together with the other evidence so far has shown this approach to be promisingly robust and suggests that real-world application is within reach.

Nonetheless, one limitation that must be overcome by the framework in order to be viable for real-time application is its low temporal precision. To close this gap between offline and online decoding, a potential solution is to look beyond envelope reconstruction towards other neural signatures of attention. Our findings in this thesis demonstrated that this avenue is certainly worth pursuing. In Chapter 3, we showed that the decoder weights themselves – that is, the model weights mapping from EEG data to the acoustic envelope – display distinct spatio-temporal characteristics as a function of whether subjects were attending to a talker, and that these characteristics were consistent across subjects. We exploited this finding to build a combined envelope-reconstruction-and-decoder-weight (env+DW) decoder and found a significant improvement in decoding performance over using envelope reconstruction accuracy alone. The inclusion of alpha power was also considered, as it has previously been shown to be hemispherically-lateralized in cocktail party tasks involving spatial attention (Wöstmann et al., 2016). However, our task was not a spatial one per se in that subjects were merely instructed to attend to a particular talker – this is arguably more akin to how people would listen in the real world – and we found that in such a scenario, alpha power lateralization in EEG was not sufficiently robust to be useful for decoding. In Chapter 5 (Section 5.3.2 and 5.3.3), we tested whether the reconstruction of relative pitch is modulated by attention (it was shown in Chapter 4 that this measure is tracked by delta-band EEG and dissociable from the acoustic envelope), and if so, whether it would be sufficiently robust for decoding. We found that the performance of this metric on its own was significantly above chance but lower than envelope reconstruction; and combining the two measures led to an improvement for some subjects.

In summary, our work on auditory attention decoding in this thesis has found the envelope reconstruction approach robust to the non-stationarity of talker location. Envelope reconstruction accuracy also remains the best metric for decoding attentional selection. Nonetheless, we found that small improvements in decoding performance can be achieved through the incorporation of other features into the decoding algorithm. Thus, we believe that looking for features beyond envelope reconstruction is an avenue worth pursuing. It could well be that if sufficiently many features with unique decoding power can be found, the aggregate of their marginal gains may lead to a large enough improvement for robust real-time performance.

6.2 The cortical processing of prosody

Spoken language conveys meaning not only via the selection of linguistic units, but also by prosody. Despite its importance, prosody has arguably received less consideration in the speech and language comprehension literature than linguistic units. Notably, one of the most influential models of speech comprehension (Hickok & Poeppel, 2000, 2004, 2007) does not explicitly account for the processing of prosodic information. As set out in Chapters 2 (Section 2.3.3) and 4 (Section 4.1), there have been studies over the years investigating the brain basis of prosody, but this research has largely been limited to anatomical localization (and in particular, testing for hemispheric lateralisation effects – see Baum & Pell, 1999; Friederici, 2011).

In the last few years, efforts have been made to incorporate prosody into the broader speech comprehension picture and to better understand the computations underpinning prosodic processing. Sammler et al. (2015) put forth a dual-stream model for prosody with a similar architecture to Hickok & Poeppel's dual-stream model for linguistic units, but with relative rightward asymmetry. Their theory was largely based on fMRI responses to isolated words with varying intonation and it focused on describing neuroanatomical pathways rather than the precise computations involved. But a notable postulation of their model is that there is a ventral ('what') pathway for prosody which has the computational role of gradually forming speaker-invariant 'prosodic Gestalts' (i.e. patterns) from speaker-dependent measures. More recently, in an effort to understand how prosody is encoded, Tang et al. (2017) recorded ECoG from subjects as they listened to synthesized sentences with varying intonation. They found that speaker-normalised relative pitch – an important acoustic percept that underlies many prosodic phenomena – is tracked by high-gamma STG activity. Tang et al.'s findings are compelling – demonstrating for the first time the direct neural encoding of speaker-invariant relative pitch and its dissociation from sites encoding absolute pitch. Nonetheless, their study utilised synthesized isolated sentences. While this is a step up from isolated words, it still lacks a lot of the richness present in continuous natural speech (e.g. the variety and liveliness of naturalistic prosody and a wider narrative context).

A major contribution of our work in Chapter 4 was to use the TRF approach – which allows for the use of stimuli that vary along multiple dimensions – to demonstrate that speaker-normalised relative pitch is tracked during listening to continuous natural speech in low-frequency (primarily lower delta-band) EEG. Importantly, we show that this tracking is dissociable from the tracking of other acoustic and phonetic features. These findings support the notion of a parallel channel for the processing of prosodic information. In a separate experiment in Chapter 5, we validated the tracking of relative pitch in the context of a cocktail-party scenario. We showed that the unique predictive power of this measure is more sensitive to attention than that of low-level acoustic (i.e. speaker-dependent) measures such as the spectrogram and its derivative. Given that relative pitch is speaker-invariant, and thus posited to be a higher-level representation in the ventral pathway of Sammler's dual-stream model, this finding is also consistent with the idea that higher-order processing areas are more affected by attention – for which there has been evidence from the visual (Kastner & Pinsk, 2004; Maunsell & Cook, 2002) and auditory (Mesgarani and Chang, 2012; Brodbeck et al., 2018; Puvvada & Simon, 2017) domains. An outstanding future question concerns how prosodic contours map onto a semantic space to form a unified percept for speech comprehension. To date, we are limited in what we know about how listeners' interpretation of utterance meaning is influenced by prosody (Roettger et al., 2019). Studies so far suggest that there is no one-to-one mapping between prosodic features and intention – it could be that listeners' weight cues based on their sensitivity to statistics in the input (Kurumada et al., 2014).

6.3 Attentional effects dissociate the acoustic and phonetic processing stages of speech comprehension

Most models of speech comprehension assume multiple levels of processing, where intermediate and increasingly abstract representations are computed in going from the auditory input to meaning. With regards to this, a longstanding controversy is whether there exists an intermediate phoneme-level representation between the mapping of the continuous sound stream to the mental lexicon. There appears to be perceptual basis for such a representation – evidence from behavioural (Liberman, Harris, Hoffman, & Griffith, 1957) and ERP studies (Näätänen, 2001) show that humans do discriminate better across than within phoneme boundaries. Nonetheless, opponents of the idea are concerned with its lack of physical basis – no acoustic cues have been found so far that generalize across instances of a particular phoneme type. It has also been argued that the aforementioned behavioural and ERP evidence could be simply due to the nature of the tasks and may not be reflective of everyday speech comprehension. In particular, they cite the observation that patients with Broca's aphasia are impaired when it comes to performing categorical perception tasks (e.g. syllable identification), but this doubly dissociates from word recognition (Lotto & Holt, 2000).

Recently, several studies have employed the use of neuroimaging and encoding/decoding methods to test whether a phonetic feature representation is tracked during listening to natural speech. Mesgarani et al. (2014) found response selectivity to phonetic features at the level of single electrodes (ECoG) in STG. Di Liberto et al. (2015) showed that phonetic features contributed uniquely to predicting low-frequency EEG over acoustic features (spectrogram and acoustic envelope) alone. And Khalighinejad et al. (2017) showed consistent time-locked responses to phonetic features. Nonetheless, a study by Daube et al. (2019) brought some of these results into question – they showed using a different analysis pipeline that the ability of phonetic features to predict EEG could be also explained by the inclusion of other acoustic feature spaces which had previously not been accounted for, such as the spectrogram derivative. Still, they acknowledged that their result cannot definitively prove that phonetic features have no explanatory power – it could well be that their pipeline was simply not sensitive enough to isolate parts of the response corresponding to phonetic features.

Given that the mapping from an acoustic to a phonetic stage involves an increase in abstraction, we had hypothesised that attentional effects might allow us to dissociate between representations at the two levels (if phonetic-level representations are indeed present). As described in the previous section (Section 6.2) and in earlier chapters of the thesis, there is evidence for the notion that higher-order processing areas are more affected by attention. In Chapter 5 of this thesis, we show that using our analysis pipeline, and in the context of a two-talker (male and female) cocktail party experiment, the phonetic feature representation of the attended talker can uniquely predict EEG above and beyond acoustic feature spaces. Crucially, we also showed that this unique predictive power was significantly modulated by attention. This pattern of results was not observed for the acoustic and feature onset representations, suggesting that phonetic features is a distinct and higher-level representation. This finding also relates to the persisting debate on whether attentional selection occurs at an early or late stage of processing. The capacity of the brain to process information from multiple streams at any given moment in time is limited, so a longstanding question has concerned the extent to which unattended streams undergo processing in cortex. While we found some evidence for the encoding of the phonetic features of the ignored talker, the effect was weak. This result considered together with previous studies (Brodbeck et al., 2018; Broderick, et al., 2018), suggests it possible that attention is categorically selective for speaker-invariant representations but attenuates lower-level acoustic (speaker-dependent) measures.

6.4 Future extensions to the work

A central theme of our work was the employment of an encoding/decoding approach to relate low-frequency EEG to different speech representations. Encoding expresses the data of each individual EEG channel as a function of stimulus features. The resultant transformation model weights are interpretable as a reflection of cognitive processes. Decoding on the other hand expresses individual stimulus features as a function of the EEG channels. The resultant model weights are not interpretable, but a better quality of fit is typically achieved. This is largely because EEG contains non-task related activity that is impossible for the encoder to predict from the stimulus features, but this noise can be spatially filtered away with the decoding approach. Nonetheless, in both cases, the model only accounts for the multivariate nature of either the stimulus or the response. This may be suboptimal particularly for applications like decoding attentional selection. Thus far, to employ the use of more than one stimulus feature for decoding auditory attention, we optimise decoders mapping from the multivariate EEG to each (continuous) feature separately. Then a classifier is trained using the resultant individual feature model reconstruction accuracies. Of late, there has been research looking at modelling approaches that jointly optimize the multivariate stimulus features and EEG response (Das, Vanthornhout, Francart, & Bertrand, 2019; de Cheveigné et al., 2018). One of these recently introduced approaches is canonical correlation analysis (CCA). CCA transforms both the stimulus and response to a space in which irrelevant variance is minimized, and has been shown to find higher correlation scores for EEG/MEG data than forward/backward modelling (de Cheveigné et al., 2018). Future work would therefore be to test whether this modelling approach can more effectively incorporate other measures such as relative pitch into the auditory attention decoding framework to improve overall decoding performance.

We explored measures of prosodic pitch in this thesis and found a signature of relative pitch in low-frequency EEG. A potential future study that builds upon this work is to test the encoding of relative pitch in populations with speech and language disorders and autism spectrum disorder (ASD). As noted in Chapter 4, many individuals with ASD display unusual or odd-sounding prosody. It would be interesting to see if our EEG index of relative pitch is sensitive to this. If found to be a robust indicator of atypical prosodic processing in small population studies, further work could then include taking practical considerations into account to facilitate widespread use. In the current work, we acquired over an hour of high-density EEG data per subject, but this may not be viable for testing in clinical environments. It would therefore be useful to determine if we could still achieve a significant effect with a reduced set of channels and less training time.

Lastly, EEG signatures of various levels of speech processing have now been found. Nonetheless, an area that remains relatively unexplored is the processing of syntax – the set of principles that govern the way words are combined to make up larger units – during listening to continuous natural speech. A recent ECoG study found evidence for phrase-structure build up (Nelson et al., 2017) in that high gamma power increased over phrases. Future work could be to identify representations/models of syntactic processing that are testable using our framework to probe how cortex encodes syntax.

6.5 Summary of contributions

The goal of the thesis was to employ an encoding/decoding approach (TRF) in conjunction with EEG to increase our overall understanding of the brain bases of speech processing and selective auditory attention and improve upon the performance of current methods for decoding selective attention. To this end, we:

a) demonstrated the robustness of an established envelope reconstruction decoding framework in a moving talker scenario, and showed that incorporating decoder model weights into the algorithm could lead to improved performance,

b) showed reliable encoding of relative pitch in delta-band EEG that is dissociable from the tracking of other acoustic and phonetic features; thus, supporting the notion of a parallel processing stream for prosody, and

c) found unique EEG predictive power for a phonetic feature representation in the context of a cocktail party study and demonstrated that it is differentially modulated by attention compared to other acoustic features; thus, supporting the notion that it is a distinct and higher-level stage of speech processing.

References

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*(23), 13367–13372. https://doi.org/10.1073/pnas.201400998

Aiken, S. J., & Picton, T. W. (2008). Human Cortical Responses to the Speech Envelope. *Ear and Hearing*, 29(2), 139. https://doi.org/10.1097/AUD.0b013e31816453dc

Aitchison, J. (2007). *The Articulate Mammal: An Introduction to Psycholinguistics*. Routledge.

Akram, S., Presacco, A., Simon, J. Z., Shamma, S. A., & Babadi, B. (2016). Robust decoding of selective auditory attention from MEG in a competing-speaker environment via state-space modeling. *NeuroImage*, *124*(Pt A), 906–917. https://doi.org/10.1016/j.neuroimage.2015.09.048

Akram, S., Simon, J. Z., & Babadi, B. (2017). Dynamic Estimation of the Auditory Temporal Response Function From MEG in Competing-Speaker Environments. *IEEE Transactions on Biomedical Engineering*, *64*(8), 1896–1905. https://doi.org/10.1109/TBME.2016.2628884

Alain, C., & Arnott, S. R. (2000). Selectively attending to auditory objects. *Frontiers in Bioscience: A Journal and Virtual Library*, *5*, D202-212.

Algazi, V. R., Duda, R. O., Thompson, D. M., & Avendano, C. (2001). The CIPIC HRTF database. *Proceedings of the 2001 IEEE Workshop on the Applications of Signal Processing to Audio and Acoustics* (Cat. No.01TH8575), 99–102. https://doi.org/10.1109/ASPAA.2001.969552

Baum, S. R., & Pell, M. D. (1999). The neural bases of prosody: Insights from lesion studies and neuroimaging. *Aphasiology*, *13*(8), 581–608. https://doi.org/10.1080/026870399401984

Bear, M. F., Connors, B. W., & Paradiso, M. A. (2007). *Neuroscience: Exploring the brain, 3rd ed.* Philadelphia, PA, US: Lippincott Williams & Wilkins Publishers.

Bednar, A., & Lalor, E. C. (2018). Neural tracking of auditory motion is reflected by delta phase and alpha power of EEG. *NeuroImage*, 181, 683–691. https://doi.org/10.1016/j.neuroimage.2018.07.054

Best, V., Ozmeral, E. J., Kopčo, N., & Shinn-Cunningham, B. G. (2008). Object continuity enhances selective auditory attention. *Proceedings of the National Academy of Sciences*, 105(35), 13174–13178. https://doi.org/10.1073/pnas.0803718105

Binder, J. R., Rao, S. M., Hammeke, T. A., Yetkin, F. Z., Jesmanowicz, A., Bandettini, P. A., … Hyde, J. S. (1994). Functional magnetic resonance imaging of human auditory cortex. *Annals of Neurology*, *35*(6), 662–672. https://doi.org/10.1002/ana.410350606

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *IFA Proceedings* 17, 97–110.

Boersma, P., & Weenink, D. (n.d.). *Praat: Doing Phonetics by Computer*, Version 6.0.20. Retrieved September 18, 2016, from http://www.fon.hum.uva.nl/praat/

Bolinger, D. (1986). *Intonation and Its Parts: Melody in Spoken English*. Stanford University Press.

Briley, P. M., Breakey, C., & Krumbholz, K. (2013). Evidence for Pitch Chroma Mapping in Human Auditory Cortex. Cerebral Cortex (New York, NY), 23(11), 2601–2610. https://doi.org/10.1093/cercor/bhs242

Bregman, A. S. (1990). *Auditory scene analysis: The perceptual organization of sound*. Cambridge, MA, US: The MIT Press.

Broadbent, D. E. (1958). The effects of noise on behaviour. *Perception and Communication.*, 81–107. https://doi.org/10.1037/10037-005

Brodbeck, C., Hong, L. E., & Simon, J. Z. (2018). Rapid Transformation from Auditory to Linguistic Representations of Continuous Speech. *Current Biology*, *28*(24), 3976-3983.e5. https://doi.org/10.1016/j.cub.2018.10.042

Broderick, M. P., Anderson, A. J., Liberto, G. M. D., Crosse, M. J., & Lalor, E. C. (2018). Electrophysiological Correlates of Semantic Dissimilarity Reflect the Comprehension of Natural, Narrative Speech. *Current Biology*, *28*(5), 803-809.e3. https://doi.org/10.1016/j.cub.2018.01.080

Carlyon, R. P. (2004). How the brain separates sounds. *Trends in Cognitive Sciences*, *8*(10), 465–471. https://doi.org/10.1016/j.tics.2004.08.008

Carlyon, R. P., Cusack, R., Foxton, J. M., & Robertson, I. H. (2001). Effects of attention and unilateral neglect on auditory stream segregation. *Journal of Experimental Psychology. Human Perception and Performance*, *27*(1), 115–127.

Carlyon, R. P., Plack, C. J., Fantini, D. A., & Cusack, R. (2003). Cross-Modal and Non-Sensory Influences on Auditory Streaming. *Perception*, *32*(11), 1393–1402. https://doi.org/10.1068/p5035

Chait, M., de Cheveigné, A., Poeppel, D., & Simon, J. Z. (2010). Neural dynamics of attending and ignoring in human auditory cortex. *Neuropsychologia*, *48*(11), 3262–3271. https://doi.org/10.1016/j.neuropsychologia.2010.07.007

Cherry, C. (1953). Some Experiments on the Recognition of Speech, with One and with Two Ears. *The Journal of the Acoustical Society of America,* 25(5), 975–979. https://doi.org/10.1121/1.1907229

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

Chertkow, H., Bub, D., Deaudon, C., & Whitehead, V. (1997). On the Status of Object Concepts in Aphasia. Brain and Language, 58(2), 203–232. https://doi.org/10.1006/brln.1997.1771

Chrabaszcz, A., Winn, M., Lin, C. Y., & Idsardi, W. J. (2014). Acoustic Cues to Perception of Word Stress by English, Mandarin, and Russian Speakers. *Journal of Speech, Language, and Hearing Research : JSLHR*, *57*(4), 1468–1479. https://doi.org/10.1044/2014_JSLHR-L-13-0279

Cohen, M. X. (2014). *Analyzing Neural Time Series Data: Theory and Practice*. MIT Press.

Cole, J., & Shattuck-Hufnagel, S. (2016). New Methods for Prosodic Transcription: Capturing Variability as a Source of Information. *Laboratory Phonology: Journal of the Association for Laboratory Phonology*, *7*(1), 8. https://doi.org/10.5334/labphon.29

Coles, M. G. H., & Rugg, M. D. (1995). Event-related brain potentials: An introduction. In *Electrophysiology of mind: Event-related brain potentials and cognition* (pp. 1–26). Oxford University Press.

Crone, N. E., Boatman, D., Gordon, B., & Hao, L. (2001). Induced electrocorticographic gamma activity during auditory perception. *Clinical Neurophysiology*, 112(4), 565–582. https://doi.org/10.1016/S1388-2457(00)00545-9

Crone, N. E., Sinai, A., & Korzeniewska, A. (2006). High-frequency gamma oscillations and human brain mapping with electrocorticography. In C. Neuper & W. Klimesch (Eds.), *Progress in Brain Research* (pp. 275–295). https://doi.org/10.1016/S0079-6123(06)59019-3

Crosse, M. J., Di Liberto, G. M., Bednar, A., & Lalor, E. C. (2016). The Multivariate Temporal Response Function (mTRF) Toolbox: A MATLAB Toolbox for

Relating Neural Signals to Continuous Stimuli. *Frontiers in Human Neuroscience*, *10*. https://doi.org/10.3389/fnhum.2016.00604

Culling, J. F., & Darwin, C. J. (1993). The role of timbre in the segregation of simultaneous voices with intersecting F0 contours. *Perception & Psychophysics*, *54*(3), 303–309. https://doi.org/10.3758/BF03205265

Cusack, R., Deeks, J., Aikman, G., & Carlyon, R. P. (2004). Effects of location, frequency region, and time course of selective attention on auditory scene analysis. *Journal of Experimental Psychology. Human Perception and Performance*, *30*(4), 643–656. https://doi.org/10.1037/0096-1523.30.4.643

Cusack, R., & Roberts, B. (2000). Effects of differences in timbre on sequential grouping. *Perception & Psychophysics*, *62*(5), 1112–1120.

Da Costa, S., van der Zwaag, W., Marques, J. P., Frackowiak, R. S. J., Clarke, S., & Saenz, M. (2011). Human primary auditory cortex follows the shape of Heschl's gyrus. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *31*(40), 14067–14075. https://doi.org/10.1523/JNEUROSCI.2000-11.2011

Darwin, C. J., & Hukin, R. W. (1997). Perceptual segregation of a harmonic from a vowel by interaural time difference and frequency proximity. *The Journal of the Acoustical Society of America*, *102*(4), 2316–2324. https://doi.org/10.1121/1.419641

Das, N., Bertrand, A., & Francart, T. (2018). EEG-based auditory attention detection: Boundary conditions for background noise and speaker positions. *Journal of Neural Engineering*, *15*(6), 066017. https://doi.org/10.1088/1741-2552/aae0a6

Das, N., Biesmans, W., Bertrand, A., & Francart, T. (2016). The effect of head-related filtering and ear-specific decoding bias on auditory attention detection. *Journal of Neural Engineering*, *13*(5), 056014. https://doi.org/10.1088/1741-2560/13/5/056014

Das, N., Vanthornhout, J., Francart, T., & Bertrand, A. (2019). Stimulus-aware spatial filtering for single-trial neural response and temporal response function estimation in high-density EEG with applications in auditory research. *NeuroImage*, 116211. https://doi.org/10.1016/j.neuroimage.2019.116211

Daube, C., Ince, R. A. A., & Gross, J. (2019). Simple Acoustic Features Can Explain Phoneme-Based Predictions of Cortical Responses to Speech. *Current Biology*. https://doi.org/10.1016/j.cub.2019.04.067

Davis, M. H., & Johnsrude, I. S. (2003). Hierarchical Processing in Spoken Language Comprehension. *Journal of Neuroscience*, 23(8), 3423–3431. https://doi.org/10.1523/JNEUROSCI.23-08-03423.2003

Drullman, R., Festen, J. M., & Plomp, R. (1994). Effect of reducing slow temporal modulations on speech reception. *The Journal of the Acoustical Society of America*, 95(5 Pt 1), 2670–2680.

Drullman, Rob, Festen, J. M., & Plomp, R. (1994). Effect of temporal envelope smearing on speech reception. *The Journal of the Acoustical Society of America*, 95(2), 1053–1064. https://doi.org/10.1121/1.408467

de Cheveigné, A., Wong, D. E., Di Liberto, G. M., Hjortkjær, J., Slaney, M., & Lalor, E. (2018). Decoding the auditory brain with canonical component analysis. *NeuroImage*, *172*, 206–216. https://doi.org/10.1016/j.neuroimage.2018.01.033

De Sanctis, P., Ritter, W., Molholm, S., Kelly, S. P., & Foxe, J. J. (2008). Auditory Scene Analysis: The interaction of stimulation rate and frequency separation on pre-attentive grouping. *The European Journal of Neuroscience*, *27*(5), 1271–1276. https://doi.org/10.1111/j.1460-9568.2008.06080.x

Delorme, A., & Makeig, S. (2004). EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Di Liberto, G. M., Wong, D., Melnik, G. A., & de Cheveigne, A. (2018). *Cortical responses to natural speech reflect probabilistic phonotactics*. https://doi.org/10.1101/359828

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*(19), 2457–2465. https://doi.org/10.1016/j.cub.2015.08.030

Ding, N., & Simon, J. Z. (2012). Emergence of neural encoding of auditory objects while listening to competing speakers. *Proceedings of the National Academy of Sciences*, *109*(29), 11854–11859. https://doi.org/10.1073/pnas.1205381109

Ding, N., & Simon, J. Z. (2014). Cortical entrainment to continuous speech: Functional roles and interpretations. *Frontiers in Human Neuroscience*, *8*. https://doi.org/10.3389/fnhum.2014.00311

Edwards, E., Soltani, M., Kim, W., Dalal, S. S., Nagarajan, S. S., Berger, M. S., & Knight, R. T. (2009). Comparison of time-frequency responses and the event-related potential to auditory speech stimuli in human cortex. *Journal of Neurophysiology*, 102(1), 377–386. https://doi.org/10.1152/jn.90954.2008

Eimas, P. D., Siqueland, E. R., Jusczyk, P., & Vigorito, J. (1971). Speech perception in infants. *Science (New York, N.Y.)*, *171*(3968), 303–306. https://doi.org/10.1126/science.171.3968.303

Engell, A. D., & McCarthy, G. (2011). The Relationship of Gamma Oscillations and Face-Specific ERPs Recorded Subdurally from Occipitotemporal Cortex. *Cerebral Cortex (New York, NY)*, 21(5), 1213–1221. https://doi.org/10.1093/cercor/bhq206

Ethofer, T., Anders, S., Erb, M., Herbert, C., Wiethoff, S., Kissler, J., Grodd, W., & Wildgruber, D. (2006). Cerebral pathways in processing of affective prosody: A dynamic causal modeling study. *NeuroImage*, *30*(2), 580–587. https://doi.org/10.1016/j.neuroimage.2005.09.059

Evans, S., McGettigan, C., Agnew, Z., Rosen, S., & Scott, S. (2016). Getting the cocktail party started: Masking effects in speech perception. *Journal of Cognitive Neuroscience*, *28*(3), 483–500. https://doi.org/10.1162/jocn_a_00913

Frey, J. N., Mainy, N., Lachaux, J.-P., Müller, N., Bertrand, O., & Weisz, N. (2014). Selective Modulation of Auditory Cortical Alpha Activity in an Audiovisual Spatial Attention Task. *Journal of Neuroscience*, 34(19), 6634–6639. https://doi.org/10.1523/JNEUROSCI.4813-13.2014

Friederici, A. D. (2011). The Brain Basis of Language Processing: From Structure to Function. *Physiological Reviews*, *91*(4), 1357–1392. https://doi.org/10.1152/physrev.00006.2011

Fritz, J. B., Elhilali, M., David, S. V., & Shamma, S. A. (2007). Auditory attention—Focusing the searchlight on sound. *Current Opinion in Neurobiology*, *17*(4), 437–455. https://doi.org/10.1016/j.conb.2007.07.011

Gandour, J., Tong, Y., Wong, D., Talavage, T., Dzemidzic, M., Xu, Y., … Lowe, M. (2004). Hemispheric roles in the perception of speech prosody. *NeuroImage*, 23(1), 344–357. https://doi.org/10.1016/j.neuroimage.2004.06.004

Gaskell, M. G., & Marslen-Wilson, W. D. (1997). Integrating Form and Meaning: A Distributed Model of Speech Perception. *Language and Cognitive Processes*, 12(5–6), 613–656. https://doi.org/10.1080/016909697386646

Gorman, K., Howell, J., & Wagner, M. (2011). Prosodylab-aligner: A tool for forced alignment of laboratory speech. *Canadian Acoustics*, 39(3), 192–193.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105*(2), 251–279.

Gray, H., & Lewis, W. H. (1918). *Anatomy of the Human Body*. Retrieved from https://books.google.com/books?id=uaQMAAAAYAAJ

Griffiths, T. D., & Warren, J. D. (2004). What is an auditory object? *Nature Reviews Neuroscience*, *5*(11), 887. https://doi.org/10.1038/nrn1538

Griffiths, T. D., Johnsrude, I., Dean, J. L., & Green, G. G. (1999). A common neural substrate for the analysis of pitch and duration pattern in segmented sound? *Neuroreport*, 10(18), 3825–3830.

Griffiths, Timothy D., Kumar, S., Sedley, W., Nourski, K. V., Kawasaki, H., Oya, H., … Howard, M. A. (2010). Direct Recordings of Pitch Responses from Human Auditory Cortex. *Current Biology*, 20(12), 1128–1132. https://doi.org/10.1016/j.cub.2010.04.044

Gussenhoven, C., & Jacobs, H. (2013). *Understanding Phonology*. Routledge.

Haegens, S., Nácher, V., Luna, R., Romo, R., & Jensen, O. (2011). α-Oscillations in the monkey sensorimotor network influence discrimination performance by rhythmical inhibition of neuronal spiking. *Proceedings of the National Academy of Sciences*, 108(48), 19377–19382. https://doi.org/10.1073/pnas.1117190108

Hall, D. A., & Plack, C. J. (2009). Pitch Processing Sites in the Human Auditory Brain. Cerebral Cortex, 19(3), 576–585. https://doi.org/10.1093/cercor/bhn108

Hamilton, L. S., & Huth, A. G. (2018). The revolution will not be controlled: Natural stimuli in speech neuroscience. *Language, Cognition and Neuroscience*, *0*(0), 1–10. https://doi.org/10.1080/23273798.2018.1499946

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. https://doi.org/10.1016/j.neuroimage.2013.10.067

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The Faculty of Language: What Is It, Who Has It, and How Did It Evolve? *Science*, *298*(5598), 1569–1579. https://doi.org/10.1126/science.298.5598.1569

Herdener, M., Esposito, F., Scheffler, K., Schneider, P., Logothetis, N. K., Uludag, K., & Kayser, C. (2013). Spatial representations of temporal and spectral sound cues in human auditory cortex. *Cortex*, *49*(10), 2822–2833. https://doi.org/10.1016/j.cortex.2013.04.003

Hickok, G. (2000). Chapter 4. Speech Perception, Conduction Aphasia, and the Functional Neuroanatomy of Language. In *Language and the Brain* (pp. 87–104). https://doi.org/10.1016/B978-012304260-6/50006-2

Hickok, G. (2013). The Functional Anatomy of Speech Processing: From Auditory Cortex to Speech Recognition and Speech Production. *In FMRI* (pp. 111–118). https://doi.org/10.1007/978-3-642-34342-1_9

Hickok, G., & Poeppel, D. (2000). Towards a functional neuroanatomy of speech perception. *Trends in Cognitive Sciences*, *4*(4), 131–138. https://doi.org/10.1016/S1364-6613(00)01463-7

Hickok, G., & Poeppel, D. (2004). Dorsal and ventral streams: A framework for understanding aspects of the functional anatomy of language. *Cognition*, *92*(1), 67–99. https://doi.org/10.1016/j.cognition.2003.10.011

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, *8*(5), 393–402. https://doi.org/10.1038/nrn2113

Hirst, D. J. (2005). Form and function in the representation of speech prosody. *Speech Communication*, *46*(3), 334–347. https://doi.org/10.1016/j.specom.2005.02.020

Holdgraf, C. R., Rieger, J. W., Micheli, C., Martin, S., Knight, R. T., & Theunissen, F. E. (2017). Encoding and Decoding Models in Cognitive Electrophysiology. *Frontiers in Systems Neuroscience*, *11*. https://doi.org/10.3389/fnsys.2017.00061

Horton, C., Srinivasan, R., & D’Zmura, M. (2014). Envelope responses in single-trial EEG indicate attended speaker in a “cocktail party.” Journal of Neural Engineering, 11(4), 046015. https://doi.org/10.1088/1741-2560/11/4/046015

Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, *195*(1), 215–243. https://doi.org/10.1113/jphysiol.1968.sp008455

Jakobson, R., & Halle, M. (1956). *Fundamentals of language*. Oxford, England: Mouton & Co.

Johnsrude, I. S., Penhune, V. B., & Zatorre, R. J. (2000). Functional specificity in the right human auditory cortex for perceiving pitch direction. *Brain*, 123(1), 155–163. https://doi.org/10.1093/brain/123.1.155

Kastner, S., & Pinsk, M. A. (2004). Visual attention as a multilevel selection process. *Cognitive, Affective, & Behavioral Neuroscience*, *4*(4), 483–500. https://doi.org/10.3758/CABN.4.4.483

Kemmerer, D. (2014). *Cognitive Neuroscience of Language*. Psychology Press.

Kerlin, J. R., Shahin, A. J., & Miller, L. M. (2010). Attentional Gain Control of Ongoing Cortical Speech Representations in a “Cocktail Party.” *Journal of Neuroscience*, *30*(2), 620–628. https://doi.org/10.1523/JNEUROSCI.3631-09.2010

Khalighinejad, B., Cruzatto da Silva, G., & Mesgarani, N. (2017). Dynamic Encoding of Acoustic Features in Neural Responses to Continuous Speech. *The Journal of Neuroscience*, *37*(8), 2176–2185. https://doi.org/10.1523/JNEUROSCI.2383-16.2017

Klatt, D. H. (1989). Review of selected models of speech perception. In *Lexical representation and process* (pp. 169–226). Cambridge, MA, US: The MIT Press.

Kotz, S. A., Meyer, M., Alter, K., Besson, M., von Cramon, D. Y., & Friederici, A. D. (2003). On the lateralization of emotional prosody: An event-related functional MR investigation. *Brain and Language,* 86(3), 366–376. https://doi.org/10.1016/S0093-934X(02)00532-1

Kreitewolf, J., Friederici, A. D., & von Kriegstein, K. (2014). Hemispheric lateralization of linguistic prosody recognition in comparison to speech and speaker recognition. *NeuroImage*, 102, 332–344. https://doi.org/10.1016/j.neuroimage.2014.07.038

Kuperberg, G. R. (2007). Neural mechanisms of language comprehension: Challenges to syntax. *Brain Research*, *1146*, 23–49. https://doi.org/10.1016/j.brainres.2006.12.063

Kurumada, C., Brown, M., Bibyk, S., Pontillo, D., & Tanenhaus, M. (2014). Rapid adaptation in online pragmatic interpretation of contrastive prosody. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 36(36). https://escholarship.org/uc/item/57h0z1p9

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*(1), 189–193. https://doi.org/10.1111/j.1460-9568.2009.07055.x

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, *54*(5), 358–368. https://doi.org/10.1037/h0044417

Lotto, A. J., & Holt, L. (2000). *The Illusion of the Phoneme*. https://doi.org/10.1184/R1/6618578.v1

Lunner, T., Rudner, M., & Rönnberg, J. (2009). Cognition and hearing aids. *Scandinavian Journal of Psychology*, *50*(5), 395–403. https://doi.org/10.1111/j.1467-9450.2009.00742.x

Luo, H., & Poeppel, D. (2007). Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. *Neuron*, *54*(6), 1001–1010. https://doi.org/10.1016/j.neuron.2007.06.004

Macken, W. J., Tremblay, S., Houghton, R. J., Nicholls, A. P., & Jones, D. M. (2003). Does auditory streaming require attention? Evidence from attentional selectivity in short-term memory. *Journal of Experimental Psychology. Human Perception and Performance*, *29*(1), 43–51.

Maddox, R. K., & Shinn-Cunningham, B. G. (2012). Influence of Task-Relevant and Task-Irrelevant Feature Continuity on Selective Auditory Attention. *JARO: Journal of the Association for Research in Otolaryngology*, *13*(1), 119–129. https://doi.org/10.1007/s10162-011-0299-7

Maffei, C., Sarubbo, S., & Jovicich, J. (2019). A Missing Connection: A Review of the Macrostructural Anatomy and Tractography of the Acoustic Radiation. *Frontiers in Neuroanatomy*, *13*. https://doi.org/10.3389/fnana.2019.00027

Marslen-Wilson, W. D. (1987). Functional parallelism in spoken word-recognition. Cognition, 25(1), 71–102. https://doi.org/10.1016/0010-0277(87)90005-9

Massaro, D. W. (1974). Perceptual units in speech recognition. *Journal of Experimental Psychology*, *102*(2), 199–208. https://doi.org/10.1037/h0035854

Maunsell, J. H. R., & Cook, E. P. (2002). The role of attention in visual processing. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *357*(1424), 1063–1072. https://doi.org/10.1098/rstb.2002.1107

Mazoyer, B. M., Tzourio, N., Frak, V., Syrota, A., Murayama, N., Levrier, O., … Mehler, J. (1993). The Cortical Representation of Speech. *Journal of Cognitive Neuroscience*, *5*(4), 467–479. https://doi.org/10.1162/jocn.1993.5.4.467

McAuliffe, Michael, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger (2017). Montreal Forced Aligner: an accurate and trainable aligner using Kaldi. Presented at the 91st Annual Meeting of the Linguistic Society of America, Austin, TX

McCann, J., & Peppé, S. (2003). Prosody in autism spectrum disorders: A critical review. *International Journal of Language & Communication Disorders*, 38(4), 325–350. https://doi.org/10.1080/1368282031000154204

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, 18(1), 1–86. https://doi.org/10.1016/0010-0285(86)90015-0

McDermott, J. H. (2009). The cocktail party problem. *Current Biology*, *19*(22), R1024–R1027. https://doi.org/10.1016/j.cub.2009.09.005

McDermott, J. H., & Simoncelli, E. P. (2011). Sound texture perception via statistics of the auditory periphery: Evidence from sound synthesis. *Neuron*, 71(5), 926–940. https://doi.org/10.1016/j.neuron.2011.06.032

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*(6), 3329–3339. https://doi.org/10.1152/jn.91128.2008

Mesgarani, Nima, & Chang, E. F. (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, *485*(7397), 233–236. https://doi.org/10.1038/nature11020

Mesgarani, Nima, Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic Feature Encoding in Human Superior Temporal Gyrus. *Science*, *343*(6174), 1006–1010. https://doi.org/10.1126/science.1245994

Meyer, M., Alter, K., & Friederici, A. (2003). Functional MR imaging exposes differential brain responses to syntax and prosody during auditory sentence comprehension. *Journal of Neurolinguistics*, 16(4), 277–300. https://doi.org/10.1016/S0911-6044(03)00026-5

Meyer, M., Alter, K., Friederici, A. D., Lohmann, G., & Cramon, D. Y. von. (2002). FMRI reveals brain regions mediating slow prosodic modulations in spoken sentences. *Human Brain Mapping*, 17(2), 73–88. https://doi.org/10.1002/hbm.10042

Mirkovic, B., Bleichner, M. G., De Vos, M., & Debener, S. (2016). Target Speaker Detection with Concealed EEG Around the Ear. *Frontiers in Neuroscience*, *10*. https://doi.org/10.3389/fnins.2016.00349

Mirkovic, B., Debener, S., Jaeger, M., & Vos, M. D. (2015). Decoding the attended speech stream with multi-channel EEG: Implications for online, daily-life applications. *Journal of Neural Engineering*, *12*(4), 046007. https://doi.org/10.1088/1741-2560/12/4/046007

Molinaro, N., & Lizarazu, M. (2018). Delta (but not theta)-band cortical entrainment involves speech-specific processing. *European Journal of Neuroscience*, 48(7), 2642–2650. https://doi.org/10.1111/ejn.13811

Møller, A. R. (2013). *Hearing: Anatomy, Physiology, and Disorders of the Auditory System*. Plural Pub. https://books.google.ie/books?id=UlEDkgAACAAJ

Moray, N. (1959). Attention in dichotic listening: Affective cues and the influence of instructions. *Quarterly Journal of Experimental Psychology*, 11(1), 56–60. https://doi.org/10.1080/17470215908416289

Näätänen, R. (2001). The perception of speech sounds by the human brain as reflected by the mismatch negativity (MMN) and its magnetic equivalent (MMNm). *Psychophysiology*, *38*(1), 1–21. https://doi.org/10.1017/s0048577201000208

Narain, C., Scott, S. K., Wise, R. J. S., Rosen, S., Leff, A., Iversen, S. D., & Matthews, P. M. (2003). Defining a left-lateralized response specific to intelligible speech using fMRI. *Cerebral Cortex (New York, N.Y.: 1991)*, *13*(12), 1362–1368. https://doi.org/10.1093/cercor/bhg083

Nelson, M. J., Karoui, I. E., Giber, K., Yang, X., Cohen, L., Koopman, H., … Dehaene, S. (2017). Neurophysiological dynamics of phrase-structure building during sentence processing. *Proceedings of the National Academy of Sciences*, *114*(18), E3669–E3678. https://doi.org/10.1073/pnas.1701590114

Norman-Haignere, S., Kanwisher, N., & McDermott, J. H. (2013). Cortical Pitch Regions in Humans Respond Primarily to Resolved Harmonics and Are Located in Specific Tonotopic Regions of Anterior Auditory Cortex. *The Journal of Neuroscience*, *33*(50), 19451–19469. https://doi.org/10.1523/JNEUROSCI.2880-13.2013

O'Sullivan, J. A., Chen, Z., Herrero, J., McKhann, G. M., Sheth, S. A., Mehta, A. D., & Mesgarani, N. (2017). Neural decoding of attentional selection in multi-speaker environments without access to clean sources. *Journal of Neural Engineering*, *14*(5), 056001. https://doi.org/10.1088/1741-2552/aa7ab4

O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., … Lalor, E. C. (2015). Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG. *Cerebral Cortex (New York, N.Y.: 1991)*, *25*(7), 1697–1706. https://doi.org/10.1093/cercor/bht355

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J.-M. (2011). FieldTrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data. *Intell. Neuroscience*, 2011, 1:1–1:9. https://doi.org/10.1155/2011/156869

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(6), 1378–1387. https://doi.org/10.1093/cercor/bhs118

Plante, E., Creusere, M., & Sabin, C. (2002). Dissociating Sentential Prosody from Sentence Processing: Activation Interacts with Task Demands. *NeuroImage*, 17(1), 401–410. https://doi.org/10.1006/nimg.2002.1182

Peretz, I. (1990). Processing of local and global musical information by unilateral brain-damaged patients. *Brain*, 113 (Pt 4), 1185–1205.

Poeppel, D. (2003). The analysis of speech in different temporal integration windows: cerebral lateralization as ‘asymmetric sampling in time.’ *Speech Communication*, 41(1), 245–255. https://doi.org/10.1016/S0167-6393(02)00107-3

Penagos, H., Melcher, J. R., & Oxenham, A. J. (2004). A Neural Representation of Pitch Salience in Nonprimary Human Auditory Cortex Revealed with Functional Magnetic Resonance Imaging. *Journal of Neuroscience*, *24*(30), 6810–6815. https://doi.org/10.1523/JNEUROSCI.0383-04.2004

Pinker, S., & Bloom, P. (1990). Natural language and natural selection. *Behavioral and Brain Sciences*, *13*(4), 707–727. https://doi.org/10.1017/S0140525X00081061

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*(9), 1497–1503. https://doi.org/10.1111/j.1460-9568.2012.08060.x

Puvvada, K. C., & Simon, J. Z. (2017). Cortical Representations of Speech in a Multitalker Auditory Scene. *Journal of Neuroscience*, *37*(38), 9189–9196. https://doi.org/10.1523/JNEUROSCI.0938-17.2017

Reddy, S., & Stanford, J. N. (2015). Toward completely automated vowel extraction: Introducing DARLA. *Linguistics Vanguard*, 1(1), 15–28. https://doi.org/10.1515/lingvan-2015-0002

Roettger, T. B., Mahrt, T., & Cole, J. (2019). Mapping prosody onto meaning – the case of information structure in American English. *Language, Cognition and Neuroscience*, 34(7), 841–860. https://doi.org/10.1080/23273798.2019.1587482

Rosenfelder, I., Fruehwald, J., Evanini, K., Seyfarth, S., Gorman, K., Prichard, H., & Yuan, J. (2014). FAVE (Forced Alignment and Vowel Extraction) Suite Version 1.1.3. https://doi.org/10.5281/zenodo.9846

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*(1278), 367–373. https://doi.org/10.1098/rstb.1992.0070

Sammler, D., Grosbras, M.-H., Anwander, A., Bestelmeyer, P. E. G., & Belin, P. (2015). Dorsal and Ventral Pathways for Prosody. *Current Biology*, *25*(23), 3079–3085. https://doi.org/10.1016/j.cub.2015.10.009

Schädler, M. R., & Kollmeier, B. (2015). Separable spectro-temporal Gabor filter bank features: Reducing the complexity of robust features for automatic speech recognition. *The Journal of the Acoustical Society of America*, *137*(4), 2047–2059. https://doi.org/10.1121/1.4916618

Scharenborg, O., Norris, D., Bosch, L., & McQueen, J. M. (2005). How should a speech recognizer work? *Cognitive Science*, *29*(6), 867–918. https://doi.org/10.1207/s15516709cog0000_37

Schnupp, J., Nelken, I., & King, A. (2011). *Auditory Neuroscience: Making Sense of Sound*. MIT Press.

Schönwiesner, M., & Zatorre, R. J. (2009). Spectro-temporal modulation transfer function of single voxels in the human auditory cortex measured with high-resolution fMRI. *Proceedings of the National Academy of Sciences*, *106*(34), 14611–14616. https://doi.org/10.1073/pnas.0907682106

Schwartz, A., McDermott, J. H., & Shinn-Cunningham, B. G. (2012). Spatial cues alone produce inaccurate sound segregation: The effect of interaural time differences. *The Journal of the Acoustical Society of America*, *132*(1), 357–368. https://doi.org/10.1121/1.4718637

Shamma, S. A., Elhilali, M., & Micheyl, C. (2011). Temporal coherence and attention in auditory scene analysis. *Trends in Neurosciences*, *34*(3), 114–123. https://doi.org/10.1016/j.tins.2010.11.002

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

Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends in Cognitive Sciences*, *12*(5), 182–186. https://doi.org/10.1016/j.tics.2008.02.003

Shinn-Cunningham, B. G., Best, V., & Lee, A. K. C. (2017). Auditory Object Formation and Selection. In J. C. Middlebrooks, J. Z. Simon, A. N. Popper, & R. R. Fay (Eds.), *The Auditory System at the Cocktail Party* (pp. 7–40). https://doi.org/10.1007/978-3-319-51662-2_2

Steinhauer, K., Alter, K., & Friederici, A. D. (1999). Brain potentials indicate immediate use of prosodic cues in natural speech processing. *Nature Neuroscience*, 2(2), 191–196. https://doi.org/10.1038/5757

Sussman, E. S., Horváth, J., Winkler, I., & Orr, M. (2007). The role of attention in the formation of auditory streams. *Perception & Psychophysics*, *69*(1), 136–152. https://doi.org/10.3758/BF03194460

Tang, C., Hamilton, L. S., & Chang, E. F. (2017). Intonational speech prosody encoding in the human auditory cortex. *Science*, *357*(6353), 797–801. https://doi.org/10.1126/science.aam8577

Teki, S., Chait, M., Kumar, S., Shamma, S., & Griffiths, T. D. (2013). Segregation of complex acoustic scenes based on temporal coherence. *ELife*, *2*. https://doi.org/10.7554/eLife.00699

Teoh, E. S., Cappelloni, M. S., & Lalor, E. C. (n.d.). Prosodic pitch processing is represented in delta-band EEG and is dissociable from the cortical tracking of other acoustic and phonetic features. *European Journal of Neuroscience*, *0*(0). https://doi.org/10.1111/ejn.14510

Teoh, E. S., & Lalor, E. (2019). EEG decoding of the target speaker in a cocktail party scenario: Considerations regarding dynamic switching of talker location. *Journal of Neural Engineering*. https://doi.org/10.1088/1741-2552/ab0cf1

Treisman, A. M. (1964). Selective attention in man. British Medical Bulletin, 20(1), 12–16.

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

Van Eyndhoven, S., Francart, T., & Bertrand, A. (2017). EEG-Informed Attended Speaker Extraction From Recorded Speech Mixtures With Application in Neuro-Steered Hearing Prostheses. *IEEE Transactions on Biomedical Engineering*, *64*(5), 1045–1056. https://doi.org/10.1109/TBME.2016.2587382

Vandenberghe, R., Nobre, A. C., & Price, C. J. (2002). The response of left temporal cortex to sentences. *Journal of Cognitive Neuroscience*, *14*(4), 550–560. https://doi.org/10.1162/08989290260045800

Wessinger, C. M., VanMeter, J., Tian, B., Van Lare, J., Pekar, J., & Rauschecker, J. P. (2001). Hierarchical organization of the human auditory cortex revealed by functional magnetic resonance imaging. *Journal of Cognitive Neuroscience*, *13*(1), 1–7.

Witteman, J., van IJzendoorn, M. H., van de Velde, D., van Heuven, V. J. J. P., & Schiller, N. O. (2011). The nature of hemispheric specialization for linguistic and emotional prosodic perception: A meta-analysis of the lesion literature. *Neuropsychologia*, 49(13), 3722–3738. https://doi.org/10.1016/j.neuropsychologia.2011.09.028

Worden, M. S., Foxe, J. J., Wang, N., & Simpson, G. V. (2000). Anticipatory biasing of visuospatial attention indexed by retinotopically specific alpha-band electroencephalography increases over occipital cortex. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *20*(6), RC63.

Wöstmann, M., Herrmann, B., Maess, B., & Obleser, J. (2016). Spatiotemporal dynamics of auditory attention synchronize with speech. *Proceedings of the National Academy of Sciences*, *113*(14), 3873–3878. https://doi.org/10.1073/pnas.1523357113

Zatorre, R. J., Evans, A. C., Meyer, E., & Gjedde, A. (1992). Lateralization of phonetic and pitch discrimination in speech processing. *Science (New York, N.Y.)*, *256*(5058), 846–849.

Zatorre, R. J., Evans, A. C., & Meyer, E. (1994). Neural mechanisms underlying melodic perception and memory for pitch. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 14(4), 1908–1919.

Zatorre, R. J, & Gandour, J. T. (2008). Neural specializations for speech and pitch: Moving beyond the dichotomies. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1493), 1087–1104. https://doi.org/10.1098/rstb.2007.2161

Zatorre, Robert J., Belin, P., & Penhune, V. B. (2002). Structure and function of auditory cortex: Music and speech. *Trends in Cognitive Sciences*, 6(1), 37–46. https://doi.org/10.1016/S1364-6613(00)01816-7