Isolating neural indices of continuous speech processing from multivariate neural data

Giovanni M. Di Liberto, B.E., M.E.

Under the supervision of Dr. Edmund C. Lalor

A dissertation submitted to the

University of Dublin, Trinity College

In fulfilment of the requirements for the degree of

Doctor of Philosophy

August, 2017

Department of Electronic and Electrical Engineering

University of Dublin, Trinity College

Declaration

I, Giovanni Di Liberto, 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.

Signed,

Giovanni M. Di Liberto

June 19th, 2017

Summary

The human ability to understand speech is underpinned by a hierarchical auditory system whose successive stages process increasingly complex attributes of the auditory input. To produce categorical speech perception, it has been suggested that this system must elicit consistent neural responses to speech tokens (e.g., syllables, phonemes) despite variations in their acoustics. This is an intermediate stage of the speech processing hierarchy, followed by lexical and semantical analyses which allow the extraction of concepts from speech sounds.

Although speech is a very important and unique aspect of humans, the cortical mechanisms that allow its efficient processing remain unclear. One of the issues is methodological, with research in auditory neuroscience often constrained to utilising artificial speech stimuli such as isolated syllables or words. Also, important aspects such as the temporal cortical dynamics have been ignored by studies that use technologies such as functional MRI. This thesis investigates the neural underpinnings of speech perception using scalp electro- and magneto-encephalography (EEG, MEG), with the aim of developing methodologies capable of isolating neural activity at different stages of the speech processing hierarchy by analysing spatio-temporal cortical dynamics in response to different features of speech.

Previous research revealed that cortical activity is entrained to the low frequency amplitude fluctuations of speech in time (i.e., amplitude envelope). However, the neural underpinnings of this phenomenon remain unclear. Here, further insights on this topic are revealed by studying the responses to continuous natural speech. In the first study (Chapter 3), participants were presented with natural speech from an audio-book while non-invasive EEG signals were recorded. This chapter demonstrates that EEG signals are sensitive to the temporal tracking of categorical phonological features of speech and provides a novel analysis framework to quantitatively investigate this phenomenon.

The study in Chapter 4 aimed to further assess this framework and, specifically, its ability to isolate cortical responses to phonetic features from the ones in response to speech acoustics. This involved implementing a perceptual pop-out paradigm that, by providing or not providing prior predictive knowledge on the upcoming stimuli, allowed for the comparison between two conditions consisting of the same stimulus but different perceived clarity. As a result, this study introduced a novel cortical measure of phoneme-

level speech processing that is modulated by the perceived clarity of the incoming stimuli. The effects of prior predictive knowledge on the cortical responses to speech are further investigated in Chapter 5, in which a similar pop-out paradigm and source-space MEG signals were used to provide new insights on the neural substrates of these effects and, specifically, on the interactions between and within selected cortical sites in temporal and frontal areas.

The other important question of this thesis regards the applicability of the novel framework in Chapter 3 in the study of language development and specific speech and language impairments. This kind of research can involve working with particular cohorts (e.g., clinical populations, infants), which may be unable to undertake long EEG experiments, and one issue relating to the framework described in the previous chapters is whether it would work with small amounts of data. Chapter 6 aims to address this by introducing an extension of the original approach that allows the reduction of the recording time down to 10 minutes. Finally, this improved approach is applied to investigate speech perception in children with dyslexia (Chapter 7) and the results support theories stating that this developmental deficit, whose symptoms involve mainly a person's reading skills, is related with language processing and, in particular, with the low frequency cortical tracking of phonological features.

Acknowledgements

This thesis is the result of four fantastic years of work and would not have been possible without the support of several people. Therefore, I would like to thank:

Ed, for his constant and thoughtful encouragement and trust. Your incredible mentorship represents a great role model both in research and life. Thank you for giving me the opportunity to travel so much, which allowed me to grow, build my research network, and make new friends. Thanks for teaching me that great ideas can arise while having a good time or a nice conversation. Also, thanks for the humiliating games of pool, Street Fighter, football, and bowling!

Richard, for his time, advice, and encouragement during these years.

John Butler, for his help at the start of my PhD and for his advice when I most needed it.

Mick, James, and Ger, who supported me greatly at the start of my PhD. Thanks Mick for the many hours of productive conversation and for his friendship, I hope we will continue collaborating on new cool studies in the future!

The current members of the LalorLab, ReillyLab, and the "tissue guys". The constant craic in the lab and on the third floor made these years very memorable.

T-Dogg Terence for his friendship, his help with figures and as a participant for my EEG experiments, and for all the trips to Il Capo!

Adam, for the hilarious and pitiless comments on papers and thesis, for the many outdoor adventures, and for his friendship.

Emily, Aisling, Michael, and Nate for the invaluable help with the thesis and for the enjoyable time exploring ideas. I look forward for more of this in the future!

Fiona, who assisted me with data collection for the study in Chapter 4.

Everybody who generously volunteered their time to participate in these studies.

Rebecca, for the fruitful collaboration and for her time while visiting York.

Varghese, Marina, and Denis, for the great opportunity to apply my work on the study of dyslexia.

The neuroscience crowd at Trinity College, for reminding me that not all neuroscience is auditory! A special thanks to David for the many football games.

Shihab, Alain, Yves, and the ENS crowd, KC, Majid, and Brad at UW, for being so welcoming during my time in Paris and Seattle and for the inspiring conversations.

Alejandro, for his invaluable friendship, for the many hours of conversation about the future, for introducing me to so many friends, and for finding me a room in a fantastic house. On this note, thanks to all my former housemates at 42 Percy Place, for making me feel at home while being on the posh side of Dublin!

My friends, for reminding me who I am and where I come from.

I miei genitori Teresa e Domenico, per il loro supporto ed incoraggiamento in questi anni trascorsi lontano da casa.

My brother Giancarlo, for being an endless source of inspiration and energy.

Pat and Ann, for being so welcoming and for making me feel part of the family.

My beautiful girlfriend Claire. For her unconditional support, for proof reading, for listening to all my waffle, and most of all for her love.

This research was kindly supported by the Irish Research Council (IRC) through a Government of Ireland Postgraduate Scholarship.

GIOVANNI DI LIBERTO

Trinity College Dublin

June 2017

Table of Contents

Publications Arising from this Thesis

Journal Publications

- **Giovanni M. Di Liberto**, O'Sullivan JA, Lalor EC, "Low frequency cortical entrainment to speech reflects phonemic level processing", *Current Biology*, **2015**, 25: 2457-2465.
- **Giovanni M. Di Liberto**, Lalor EC, "Indexing cortical entrainment to natural speech at the phonemic level: Methodological considerations for applied research", *Hearing Research*, **2017**, http://doi.org/10.1016/j.heares.2017.02.015
- **Giovanni M. Di Liberto**, Michael J. Crosse, Edmund C. Lalor, "Cortical measures of phoneme-level speech encoding correlate with the perceived clarity of natural speech", *in review*.
- **Giovanni M. Di Liberto**, Edmund C. Lalor, Rebecca E. Millman, "Causal cortical dynamics of a predictive enhancement of speech intelligibility", *in review*.
- **Giovanni M. Di Liberto**, Varghese Peter, Marina Kalashnikova, Denis Burnham, Edmund C. Lalor, "Deficits in right hemisphere encoding of natural speech correlate with psychometric measures of dyslexia", *in preparation*.

Conference Publications

- **Giovanni M. Di Liberto**, Edmund C. Lalor, Isolating neural indices of continuous speech processing at the phonetic level. *Advances in Experimental Medicine and Biology, **2016**, 894: 337-345.*

Other Related Journal Publications

- Michael J. Crosse, **Giovanni M. Di Liberto**, Edmund C. Lalor, "Eye Can Hear Clearly Now: Mechanisms Underlying Multisensory Integration in Peri-Threshold Speech-in-Noise", *Journal of Neuroscience, **2016**,* 36 (38), 9888-9895.

- Michael J. Crosse, **Giovanni M. Di Liberto**, Adam Bednar, Edmund C. Lalor, "The multivariate temporal response function (mTRF) toolbox: a MATLAB toolbox for relating neural signals to continuous stimuli", *Frontiers in Human Neuroscience*, **2016**, 10:604.
- Aisling E. O'Sullivan, Michael J. Crosse, **Giovanni M. Di Liberto**, Edmund C. Lalor, "Visual Cortical Entrainment to Motion and Categorical Speech Features during Silent Lipreading", *Frontiers in Human Neuroscience*, **2016**, 10:679.
- Mark Hasegawa-Johnson, Adrian KC Lee, Edmund C. Lalor, Preethi Jyothi, Dan McCloy, Majid Mirbagheri, **Giovanni M. Di Liberto**, Amit Das, Brad Ekin, Chunxi Liu, Vimal Manohar, Hao Tang, Nancy Chen, Paul Hager, Tyler Kekona, and Rose Sloan, "ASR for Under-Resourced Languages from Probabilistic Transcription", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **2017**, 25 (1), 50-63.
- Marina Kalashnikova, Varghese Peter, **Giovanni M. Di Liberto**, Edmund C. Lalor, Denis Burnham, "Infant directed speech facilitates neural entrainment to speech in seven-month- old infants", *in preparation*

Poster and Oral presentations

- **Poster presentation**, Neuroscience, SfN's 46th annual meeting, San Diego, USA, November 2016, "Investigating the effect of perceptual enhancement on the cortical representation of speech using source space analysis with MEG"
- **Poster presentation**, Advances and Perspectives in Auditory Neuroscience (APAN), San Diego, USA, November 2016, "Investigating the cortical encoding of phonological features of continuous speech in dyslexia"

- **Oral presentation**, Synaesthesia and Cross-Modal Perception, Dublin, April 2016, “Modelling the cortical representation of auditory and visual speech features in low frequency EEG”
- **Poster presentation**, Computational and System Neuroscience (Cosyne), Salt Lake City, USA, February 2016, “A model-based EEG approach for investigating the hierarchical nature of continuous speech processing”
- **Oral presentation**, Cognitive Science Arena, Brixen, Italy, February 2016, “Isolating neural indices of continuous speech processing at the phonetic level”
- **Poster presentation**, Workshop on Auditory Neuroscience, Cognition and Modelling, QMUL, London, UK, February 2016, “A model-based EEG approach for investigating the hierarchical nature of continuous speech processing”
- **Poster presentation**, Neuroscience, SfN's 45th annual meeting, Chicago, USA, October 2015, “Isolating neural indices of continuous speech processing at the phonetic level”
- **Poster presentation**, Advances and Perspectives in Auditory Neuroscience (APAN), Chicago, USA, October 2015, “Isolating neural indices of continuous speech processing at the phonetic level”
- **Oral presentation**, 17th International Symposium on Hearing, Groningen, Netherlands, June 2015, “Isolating neural indices of continuous speech processing at the phonetic level”
- **Poster presentation**, The Auditory Model Workshop, Oldenburg, June 2015, “Indexing neural responses to continuous speech at the phonetic level”
- **Oral presentation**, École Normale Supérieure, Paris, France, May 2015
- **Poster presentation**, 3rd Annual Trinity Centre for Bioengineering Symposium, TCD, Dublin, Ireland, December 2012, “Pushing beyond the envelope”.

- **Poster presentation**, Auditory Cortex, Magdeburg, Germany, September 2014, "Pushing beyond the envelope"
- **Poster presentation**, Young Neuroscientists Symposium, Dublin, Ireland, September 2014, "Pushing beyond the envelope"

Glossary of Acronyms

ABR	Auditory Brainstem Response
AEP/F	Auditory Evoked Potential/Field
AESPA	Auditory Evoked Spread Spectrum Analysis
BA	Brodmann Areas
BCI	Brain-Computer Interface
BOLD	Blood Oxygenation Level Dependent signal
CA	Age matched Control group
CN	Cochlear Nucleus
DX	Dyslexia
ECoG	ElectroCorticoGraphy
EEG	ElectroEncephaloGraphy
EMG	ElectroMyoGraphy
EOG	ElectroOculoGraphy
ERP/F	Event-Related Potentials/Field
FDR	False Discovery Rate
fMRI	functional Magnetic Resonance Imaging
HG	Heschl's gyrus
IC	Inferior Colliculus
IFG	Inferior Frontal Gyrus
ITS	Inferior Temporal Sulcus
LLR	Long Latency Response
LTI	Linear Time-Invariant
MDS	Multi-Dimensional Scaling
MEG	MagnetoEncephaloGraphy
MGN	Medial Geniculate Nucleus
MLR	Middle Latency Response
MRI	Magnetic Resonance Imaging
MTG	Middle Temporal Gyrus
mTRF	multivariate Temporal Response Function
PP	Planum Polare
PT	Planum Temporale
RL	Reading-Level matched control group
SF	Sylvian Fissure
SI	System Identification
SLI	Specific Language Impairment
SNR	Signal-to-Noise Ratio
SOC	Superior Olivary Complex
STG	Superior Temporal Gyrus
STRF	Spectro-Temporal Response Function
STS	Superior Temporal Sulcus
TRF	Temporal Response Function

List of Figures and Tables

Chapter 1. Introduction

Spoken language is the form of communication that most markedly distinguishes humans from other species. Humans are able to understand speech in real-time despite challenges such as noisy environments, competing speakers, co-articulation effects, and the variability of speaker accents and voices. The seeming ease of this process makes us oblivious of the incredibly complex neural operations required to parse real-world acoustic signals into their corresponding meanings (Naatanen and Winkler, 1999).

While the precise neurophysiological mechanisms and neuroanatomic infrastructure underpinning this ability are not well understood (Poeppel, 2014), robust speech perception has been proposed to be the product of a hierarchical auditory processing system whose successive stages process increasingly complex attributes of the audio input (Chang et al., 2010; Okada et al., 2010; Peelle et al., 2010). In this context, it has been suggested that, while earlier areas of the auditory system undoubtedly respond to acoustic differences in speech tokens, later areas must exhibit consistent neural responses to those tokens in order to produce a categorical perception of speech sounds and words. This framework, in which potentially discriminable speech sounds are assigned to functionally equivalent classes, suggests that speech is "special" to our brains as it is processed differently from other non-speech sounds (Liberman et al., 1967; Liberman, 1970; Stevens and House, 1972; Nourski, 2017). In particular, research in the field of speech perception has focused on the categorical perception of the smallest linguistic units that can change the meaning of a speech message, called phonemes (e.g., /c/ and /f/ as in *car* vs *far*) (Chomsky and Miller, 1968; Casserly and Pisoni, 2010). The mapping of acoustic sounds into phonemes is considered an essential step for speech comprehension. However, unveiling the cortical mechanisms that underpin such a neural process is an incredibly complex task (Casserly and Pisoni, 2010).

The human brain is composed of about 100 billion neurons interconnected with over 100 trillion synapses (Williams and Herrup, 1988; Herculano-Houzel, 2009). Today's technology is far from being able to capture such an incredible level of complexity. Specifically, there is no machine capable of measuring brain dynamics non-invasively with both high temporal and spatial resolution. For this reason, different fields in neuroscience have specialised in the study of brain mechanisms at distinct levels of detail. In the context of speech, functional magnetic resonance imaging (fMRI) (Obleser et al., 2007; DeWitt and Rauschecker, 2012), non-human primate electrophysiology (Rauschecker and Scott, 2009), and electrocorticography (ECoG) (Chang et al., 2010; Zion Golumbic et al., 2013b; Mesgarani et al., 2014; Nourski, 2017) have all made important contributions to our understanding of hierarchical speech encoding in the brain. However, all of these methods have their shortcomings: fMRI has relatively poor temporal resolution, which is a strong limitation given that speech has very fast dynamics; studies based on primates have to deal with fundamental differences with humans (e.g., primates do not speak); ECoG studies are limited to patient groups with severe cases of epilepsy.

Recent developments in electro- and magneto-encephalography (EEG/MEG) techniques may offer important opportunities for further progress. These approaches allow for the macroscopic, non-invasive study of brain patterns with high temporal resolution and have been used for years to study the processing of speech units such as discrete syllables (i.e., syllables presented in isolation) (Salmelin, 2007). Furthermore, it has been shown that both EEG and MEG index the cortical tracking of the low frequency amplitude envelope of natural speech (Aiken and Picton, 2008b; Lalor and Foxe, 2010; Ding and Simon, 2012a). This has proved useful for investigating the mechanisms underlying speech processing (Luo and Poeppel, 2007), how such processing is affected by attention (Kerlin et al., 2010; Ding and Simon, 2012a; Power et al., 2012), and how audio and visual speech interact (Luo et al., 2010; Zion Golumbic et al., 2013a). However, it has not yet been established to what extent these EEG/MEG indices reflect higher-level speech-specific processing versus lower-level processing of the spectro-temporal/acoustic stimulus dynamics (Ding and Simon, 2014).

The main goal of this thesis is the identification of an experimental and analytic approach to isolate quantitative indices of phonological processing based on measures of cortical entrainment to natural speech features using non-invasive technologies. The work presented in this manuscript is the result of collaborative, multi-disciplinary research that,

from a core of engineering methodologies, provides new insights in neighbouring fields such as neuroscience, linguistics, psychology, and clinical research.

1.1 Background

The human brain allows for effortless, robust, and real-time processing of speech. As mentioned above, speech comprehension is a hard task that requires a very complex neural network. One example that may help the reader to appreciate the importance and complexity of this task can be found in the context of speech recognition software, in the field of computer science and engineering. While we may consider speech comprehension as a straightforward process that does not require out-of-the-ordinary intelligence, that same process is one which the leading companies in the information technology field are striving to emulate in the form of speech recognition software and represents a multi-billion dollar industry. Yet, the performances of such software remain far from being comparable to human-level speech recognition.

Successful speech comprehension feels so natural and easy to do that it can be readily taken for granted. In fact, before the availability of modern signal processing technology, researchers believed that the corresponding neural processes were fairly uncomplicated and straightforward (Casserly and Pisoni, 2010). Of course, this is not the case, which is the reason for the present research work. A brilliant analogy of the modern challenges of research in speech perception was proposed by Prof. Nima Mesgarani during a talk at the 20th Jelinek Workshop in Speech and Language Technology (Seattle, USA, 2015, here readapted to the domain of neurophysiology). Let us consider a lake populated with fish that swim in various directions (**Figure 1.1A**) and that we want to study the movements of these fish. Now, all we have are measurements of wave height at different locations of the lake. In neurophysiology, measurements provided by EEG, MEG, or ECoG can be thought of as the time-series of such height values, while the fish movement corresponds to the neural activity. This type of measurement will allow the study of shoals of fish, however the movement of a single fish cannot be detected. This problem, which already seems very hard, is further complicated by various types of noise that is often larger than the signal of interest, e.g., the wave movement induced by a speed-boat (**Figure 1.1B**). The sections to follow will extensively address traditional and novel approaches to "tackle" this type of issue.

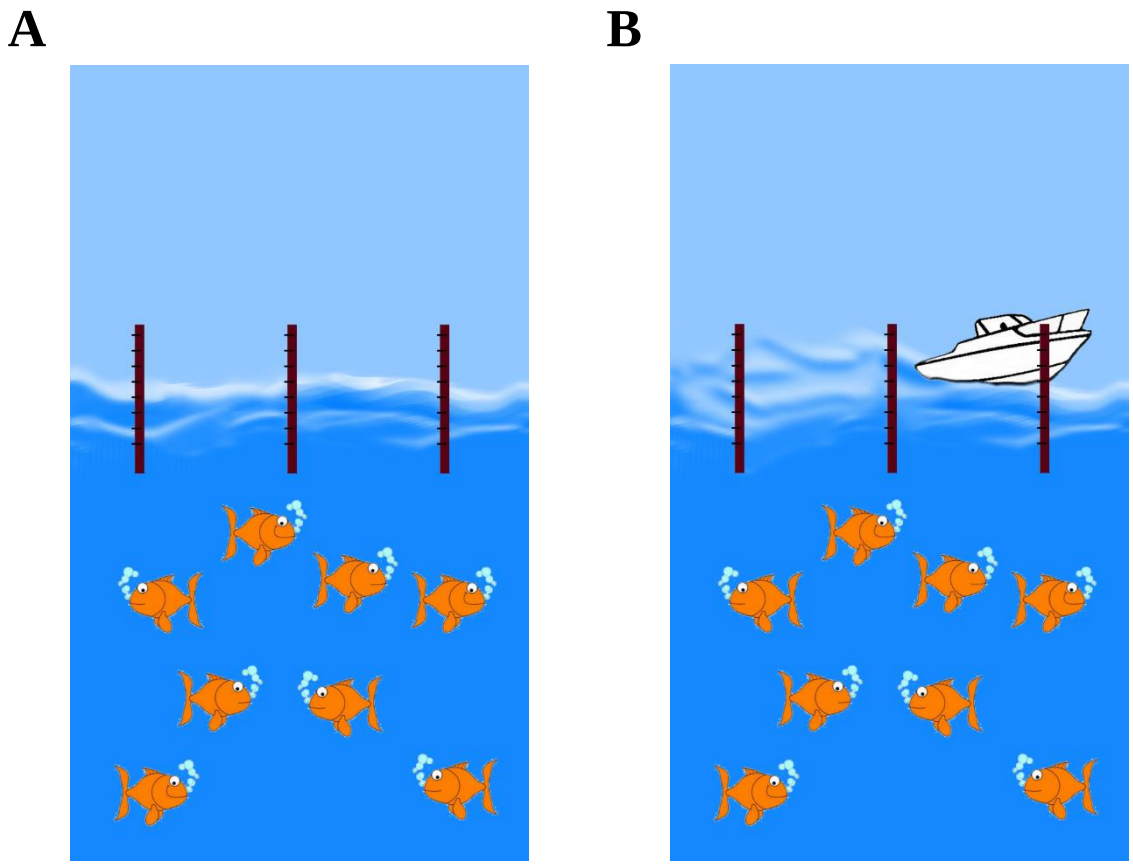


Figure 1.1: The challenge of research in speech perception.
If we think of a brain as a lake, where each neuron is a fish, neurophysiological measurements (e.g., EEG) can be thought of as the wave heights at various locations. Such measurements can be used to tell us something about the movement of shoals of fish, while deriving dynamics of each single fish is an ill-posed problem.

Despite these challenges, it is important to pursue this type of research because of the major role that successful speech comprehension plays in education and social development. This role becomes indisputable when considering the negative effects of impairments in speech and language functions, which range from the substantially higher rates of depression and anxiety, school dropout, delinquency, to a lower chance of future employment (Sabornie, 1994; Wiener and Schneider, 2002; McNulty, 2003; Daniel et al., 2006; Baker and Ireland, 2007; Brooks, 2014). Indeed, the causes and symptoms are of various types and our inability to investigate the brain dynamics of this process in sufficient detail makes it very difficult to fully understand and diagnose such deficits.

Impairment in speech and language functions can involve difficulties in comprehension and/or use of spoken, written and/or other symbolic systems. These deficits can be grouped into (American Speech-Language-Hearing Association 1993):

- *Hearing disorders* are limitations in the sensitivity to physical acoustic sounds, due to damage of the peripheral organs responsible for the perception

of sounds. Individuals with hearing impairment can be deaf (they cannot hear sounds) or hard of hearing (their primary sensory input for communication is still auditory, although sound perception is impaired).

- *Language disorders* involve the form of language (phonology, morphology, syntax), the semantic content of language, or the function of language (functional and socially appropriate communication);
- *Speech disorders* refer to a deficit in the articulation of speech sounds (mistakes, omissions, or additions), fluency (e.g., a disrupted flow of speaking, repetitions, atypical rate), or voice (atypical vocal loudness, pitch, duration, and more generally vocal quality).

Deficits in speech perception that cannot be attributed to any of these groups can be classified as central auditory processing disorders, which may involve, for example, difficult hearing in noisy environments, lack of attention during listening, or lack of auditory memory. It is important to highlight that this grouping is not clear-cut. In fact, the presence of one type of impairment does not exclude other symptoms. The ability to quantify speech processing at specific levels of the cortical hierarchy may be crucial to better understand the neural causes of these disorders, providing a means to better distinguish different deficits and to provide potential biomarkers for diagnosis. Furthermore, such neural indices could provide new insights on the cortical mechanisms of speech perception in healthy people, as well as their development from infancy to adulthood, and their decline from adult to old age.

This thesis is centred on speech perception, however this field of research is intertwined with non-verbal receptive communication. In fact, higher levels of the speech and language processing hierarchy are thought to process modality-independent information, which may be reached and shared by different pathways originating from distinct sensory inputs (Poeppel, 2006; Simanova et al., 2014; Handjaras et al., 2016). For example, one can consider the case of developmental dyslexia, a learning disorder that affects 5-10% of school aged children whose symptoms and impact persist into adulthood (Vellutino et al., 2004). Dyslexia is defined by difficulties in acquiring reading despite otherwise normal sensory and intellectual functioning. However, current theories suggest that the root causes of dyslexia are related to a deficit in the processing of phonological units (Goswami, 2011; Richlan, 2012; Goswami and Leong, 2013; Lehongre et al., 2013; Clark et al., 2014; Goswami, 2015). In fact, this processing stage is not exclusive to verbal speech, as it involves, for example, the association of written letters (*graphemes*) or lip

movements (*visemes*) with sounds. The ability to investigate speech perception at the level of phonemes could provide a new view into deficits in phonological processing for both verbal and non-verbal receptive (and expressive) communication.

1.2 Research Goals and Collaborations

The overarching goal of this thesis is to identify a novel methodology to extract quantitative indices of the cortical processing of speech at the level of phonemes using non-invasive EEG. Firstly, such a methodology is described and used to test the hypothesis that the cortical tracking of speech measured with EEG reflects speech processing at both the acoustic and phonological levels. Secondly, this thesis takes a new look at the neural underpinnings of speech perception, with a focus on the mechanism of integration of prior information and sensory input. In particular, the ability to investigate phonological processing with high temporal resolution provides an opportunity to clarify the precise dynamics and interactions between distinct hierarchical levels during speech perception. Finally, the present research work aims to define a framework to isolate indices of speech processing at the acoustic and phonological levels that is applicable in clinical and other cohorts of interest as a research tool and potentially as a diagnostic biomarker.

This research work involved international collaborations that complemented the existing infrastructure at Trinity College. Firstly, the collaboration with Dr. Rebecca E. Millman (University of York, now at University of Manchester) provided an opportunity to investigate the cortical dynamics of speech perception with high spatial and temporal resolution using MEG, a technology that is currently unavailable in the Republic of Ireland (Chapter 5). Furthermore, the interaction with Dr. Varghese Peter and Prof. Denis Burnham (MARCS institute for brain, behavior, and development; Western Sydney University) allowed for the first clinical application of this methodology in cohorts with developmental dyslexia (Chapter 7).

1.3 Thesis outline

In **Chapter 2**, the neural anatomy of speech processing is described, with a particular focus on the functional organisation in the human cortex. The cortical mechanisms that allow the integration of prior knowledge and information from other sensory modalities

into the perception of speech are discussed. As this topic remains hotly debated, two chapters of this thesis are focused on such neural mechanisms (Chapters 4 and 5). The subsequent section describes non-invasive technologies to investigate speech perception in humans. Given the importance of time-domain analysis in speech perception, a particular focus is given to electro- and magnetoencephalography (EEG and MEG respectively), which provide direct non-invasive measures of brain activity with high temporal resolution. Finally, this section includes a description of recent analysis approaches that allow the investigation of speech processing using naturalistic stimuli such as continuous speech.

Chapter 3 introduces a novel analysis framework that enables the isolation of cortical responses to speech at the level of phonemes using non-invasive EEG and an experiment with natural speech. This study provides the first evidence that EEG is sensitive to the temporal tracking of categorical phonological features of natural speech and shows sensitivity to distinct features at the level of phonemes e.g., vowels and consonants, nasal and fricative.

The availability of such a framework enables the study of a variety of questions. **Chapter 4** investigates how quantitative measures of cortical tracking of acoustic and phonetic speech features are affected by intelligibility and prior knowledge. This study demonstrates that indices of phoneme-level speech processing increase in magnitude with perceived intelligibility. Also, these measures are sensitive to the effect of prior knowledge, which is discussed in the context of current theories.

Chapter 5 builds on Chapter 4 by examining the cortical mechanisms that allow the integration of prior knowledge with sensory input using technology with higher spatial resolution. In particular, MEG (co-registered with MRI) was used to further investigate these mechanisms of integration in source-space by using source reconstruction of MEG signals in cortical locations of interest.

Another research question of this thesis is whether the novel framework introduced in Chapter 3 could find applicability in the study of speech perception in particular cohorts, e.g., clinical populations, infants. One issue is the long experimental time required for this analysis (> 1 hour). For instance, clinical assessment may involve a long battery of tests, which imposes strict time constraints on the duration of each specific test. **Chapter 6** provides a solution to this issue by introducing an extension of the analysis framework. This improved approach enables the extraction of indices of cortical tracking at the level of phonemes using only 10 minutes of recording time.

Chapter 7 describes the first application of this analysis framework in the study of speech perception in particular cohorts. Specifically, measures of cortical tracking of acoustic and phonetic features are used to investigate speech perception in developmental dyslexia in children aged between 6 and 12 years. This study reveals causal effects of dyslexia that support theories stating that this developmental deficit is related with language processing and, specifically, with the low frequency cortical tracking of phonological speech features.

Finally, **Chapter 8** presents a discussion of the main findings of the research carried out in this thesis and outlines other current and future work that builds upon the foundations of this work.

Chapter 2. The neurophysiology of speech perception

This chapter provides a review of the literature relevant to this thesis and is divided into three sections. The first section describes speech and summarises the current understanding of the functional neuroanatomy of speech comprehension, with a focus on phoneme-level processing. The following section focuses on current views on the mechanisms that allow the active integration of prior information to facilitate speech comprehension. The third section presents an overview of the methodologies to study sound and speech perception, with a focus on non-invasive technologies and continuous acoustic stimuli.

2.1 The functional anatomy of speech perception

Speech is the verbal means of communicating and consists of articulation, voice, and fluency. It can be decomposed into smaller units such as sentences, phrases, words, and syllables. In phonology, the most basic unit of sound is called 'phoneme' and it can be described as the smallest contrastive linguistic unit which may differentiate one word from another. Separately, *language* refers to the socially shared rules that allow speech to be an efficient communication method. For instance, it defines the meaning of words, how to build new words by combining different suffixes or prefixes, and how to put words together. The combination of sensory organs and neuronal structures that allow us to perceive and localise sounds, and to interpret speech sounds into meaningful messages is called the *human auditory system*. This dissertation presents research on the underpinnings of receptive speech processing specifically in the human cerebral cortex.

2.1.1 The auditory system

In its physical form, speech is a pressure waveform that travels from a speaking person to one or more listeners. This vibration is mechanically transformed into electrical signals by the human *peripheral auditory system*. The signal from the two ears are combined in the *central auditory system*, which is thought to be crucial in the extraction of early acoustic features, such as binaural cues for sound localisation and pitch. The resulting electrical signal is then transported to the *auditory cortex* which, in the context of speech comprehension, is responsible for the identification of auditory stimuli such as speech, rather than any other sound, and for the extraction of its corresponding meaning. The mechanisms that characterise the ascending auditory pathway, its evolution from animals to humans, and its impairment in case of trauma or disease have been topics of intense investigation (Salmelin, 2007; Rauschecker and Scott, 2009; Ding and Simon, 2014) (**Figure 2.1**).

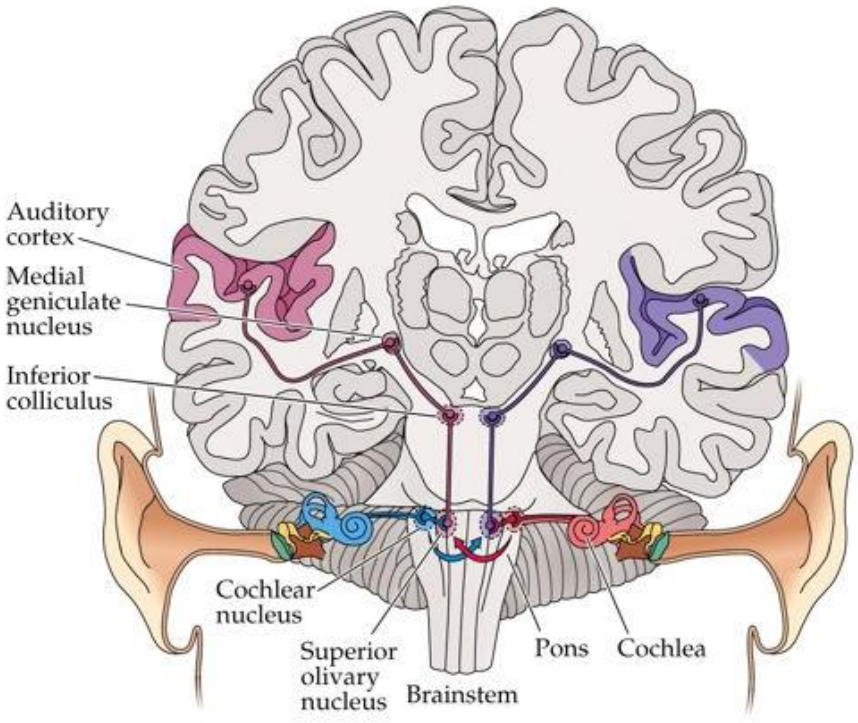


Figure 2.1: The ascending auditory pathway.
Illustration from Purves et al. (2008).

Peripheral auditory system

Sound waves arrive at the outer ear, propagate through the ear canal, and hit the *tympanic membrane* (i.e., ear drum), causing it to vibrate. This thin and conical component marks the beginning of the middle ear, which is an air-filled cavity containing three auditory

ossicles (*malleus*, *incus*, and *stapes*). Because of this structure, vibrations of the tympanic membrane propagate through the middle ear and reach the inner ear, which represents a direct interface with the nervous system (**Figure 2.2A**). These vibrations reach the cochlea, a small coiled cavity with tonotopic organisation (frequency-to-place mapping), which performs frequency analysis of the incoming sound wave by splitting it up into multiple frequency-bands (Yang et al., 1992).

The section cochlea contains tiny mechanosensory cells, known as hair cells. The vibration that reaches the cochlea induces the movement to these hair cells, which are connected to spiral ganglion neurons from the auditory nerve. Therefore, that movement of hair cells is transformed into electrical signals which, on a broader picture, allows for the transformation of an incoming sound wave into electrical signals.

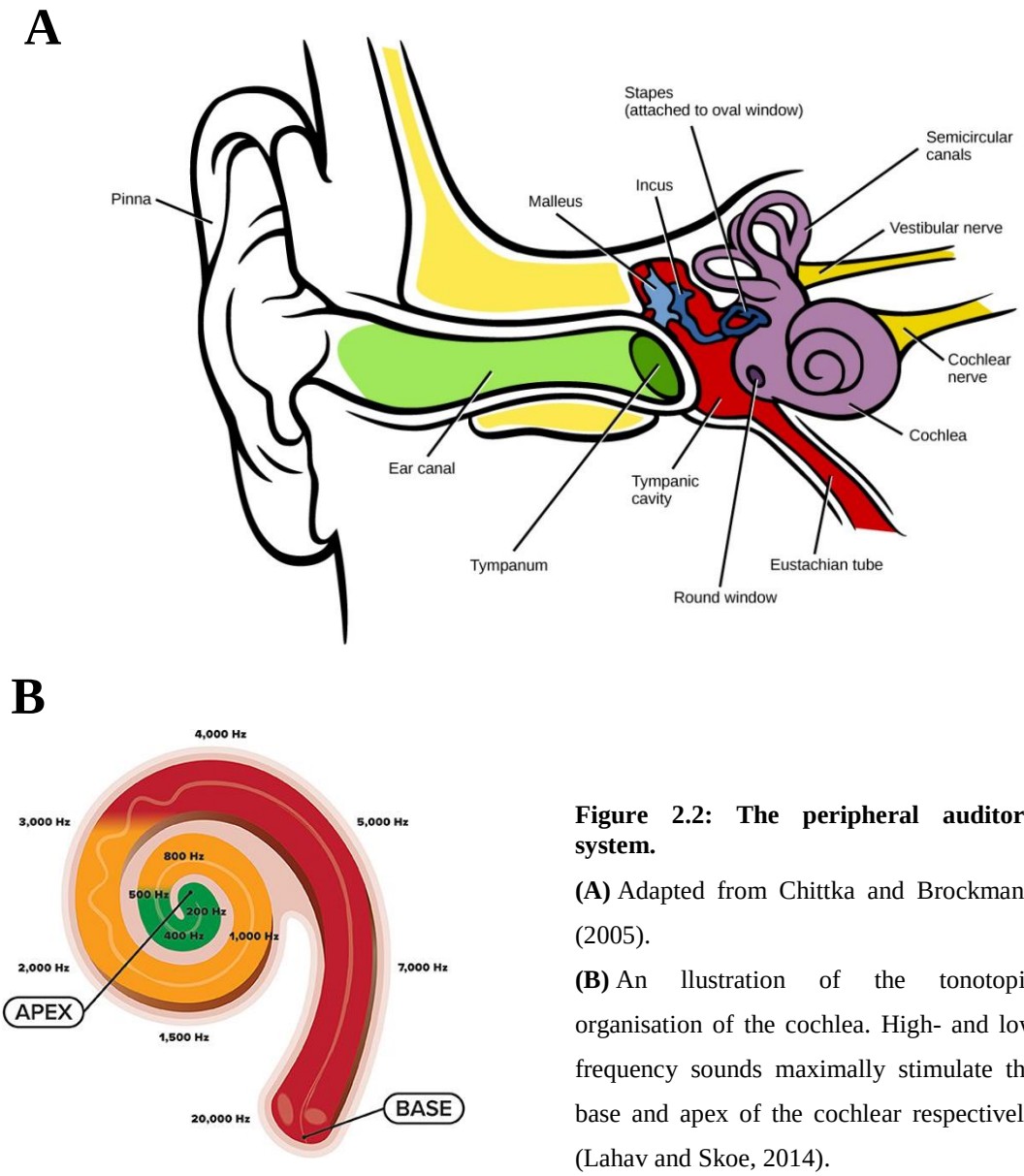


Figure 2.2: The peripheral auditory system.

(A) Adapted from Chittka and Brockmann (2005).

(B) An llustration of the tonotopic organisation of the cochlea. High- and low frequency sounds maximally stimulate the base and apex of the cochlear respectively (Lahav and Skoe, 2014).

The tonotopy of the cochlea constitutes the first stage of sound processing, which results in the filtering of the auditory signal into logarithmically spaced frequency bands (**Figure 2.2B**). This tonotopic structure is then preserved through the central auditory system and primary auditory cortex.

Central auditory system

The structures that transport and process auditory information to the human cortex after its transformation into electrical signals are collectively called the central auditory system. The inner ear communicates with the central auditory system via the auditory nerve. The *cochlear nucleus* (CN) is a structure situated in the brainstem that extracts information on the firing rate of the auditory nerve fibres and performs nonlinear spectral and spatial analysis (Purves et al., 2008). Auditory information is then transferred laterally to the *superior olivary complex* (SOC) where, for the first time, information from both ears converge and binaural cues can be derived. As a result, sound localisation in the azimuthal plane (left vs right) can be performed at this stage.

The following stage of the auditory pathway is the *inferior colliculus* (IC), which is considered to be involved in, among the other things, the integration and routing of multi-modal sensory perception and in the processing of amplitude modulation features crucial for pitch detection (Joris et al., 2004). Furthermore, it is thought to play a central role in the extraction of sound features that are critical for sound perception (Ehret and Romand, 1997). Finally, the signals propagate through the *medial geniculate nucleus* (MGN) of the thalamus, which connects IC and the primary auditory cortex. Please refer to the books by Schnupp et al. (2011) and by Clark (2006) for further reading on the structures of the auditory system described above.

Auditory cortex

The signal processed in MGN is transmitted to the auditory cortex, an area located bilaterally on the supratemporal plane that comprises the superior two-thirds of the *superior temporal gyrus* (STG) (Celesia, 1976; Galaburda and Sanides, 1980; Rivier and Clarke, 1997). To date, because of both technical and neuroscientific limitations, there is no dominant anatomical parcellation scheme of human auditory cortical areas that is generally accepted and routinely used across laboratories (Moerel et al., 2014). Similar to the monkey (which can be studied in more detail with invasive studies), the human auditory cortex has been suggested to be hierarchically organised with a *core* of primary

auditory areas that receive ascending signals from the thalamus, and is surrounded by non-primary regions named *belt* and *parabelt* (Hackett et al., 1998; Morosan et al., 2001). However, at the finer level of area definition, there is large variability among the various reports. For this reason, the following analysis is limited to a macroscopic scale.

In the anterior to posterior direction, the human auditory cortex can be divided into three regions (Kim et al., 2000): *Planum polare* (PP), *transverse temporal gyrus* or *Heschl's gyrus* (HG), and *planum temporale* (PT). HG, which is hidden in the depth of the *Sylvian fissure* (SF), is thought to be evolutionary new as it was found only in a subset of chimpanzee brains (Hackett et al., 2001), but not in the macaque monkey (however, see Baumann et al., 2013). Interpretations vary for a placement of the human auditory core along HG, across HG, and everything in between (Baumann et al., 2013). Despite this debate, it is commonly agreed that the core is located in a tonotopic area of the auditory cortex with frequency selectivity organised in a high-low-high gradient (Baumann et al., 2013; Saenz and Langers, 2014). Please refer to Moerel, De Martino, and Formisano (2014) for a more in-depth discussion on the structures and functions of the human auditory cortex.

The brain network responsible for the processing of complex sounds, such as speech, involves areas beyond the STG. In particular, other cortical sites of interest are placed in *superior temporal sulcus* (STS), *middle temporal gyrus* (MTG), and *inferior temporal sulcus* (ITS), and extend to areas of other cortical lobes, such as the *inferior frontal gyrus* (IFG). The following section describes the current views on the functional structure, roles, and interactions within such a network.

2.1.2 Hierarchical organisation of the auditory system in the human cortex

The neural organisation of speech perception in the human cortex has been surprisingly difficult to characterise, even in gross anatomical terms (Hickok and Poeppel, 2007). The hypotheses on this topic before the introduction of modern neuroimaging techniques were mainly based on observations of patients with speech disorders due to lesions in parts of the cortex. For instance, patients with speech comprehension disorders were typically affected by lesions in the left superior temporal gyrus (STG) (Hickok and Small, 2015). Therefore, this area was thought to be critical for speech perception. However, later studies revealed that the destruction of the left STG causes deficits in speech production,

not in speech comprehension, which confirmed that multiple regions participate in the speech comprehension process (Damasio and Damasio, 1980).

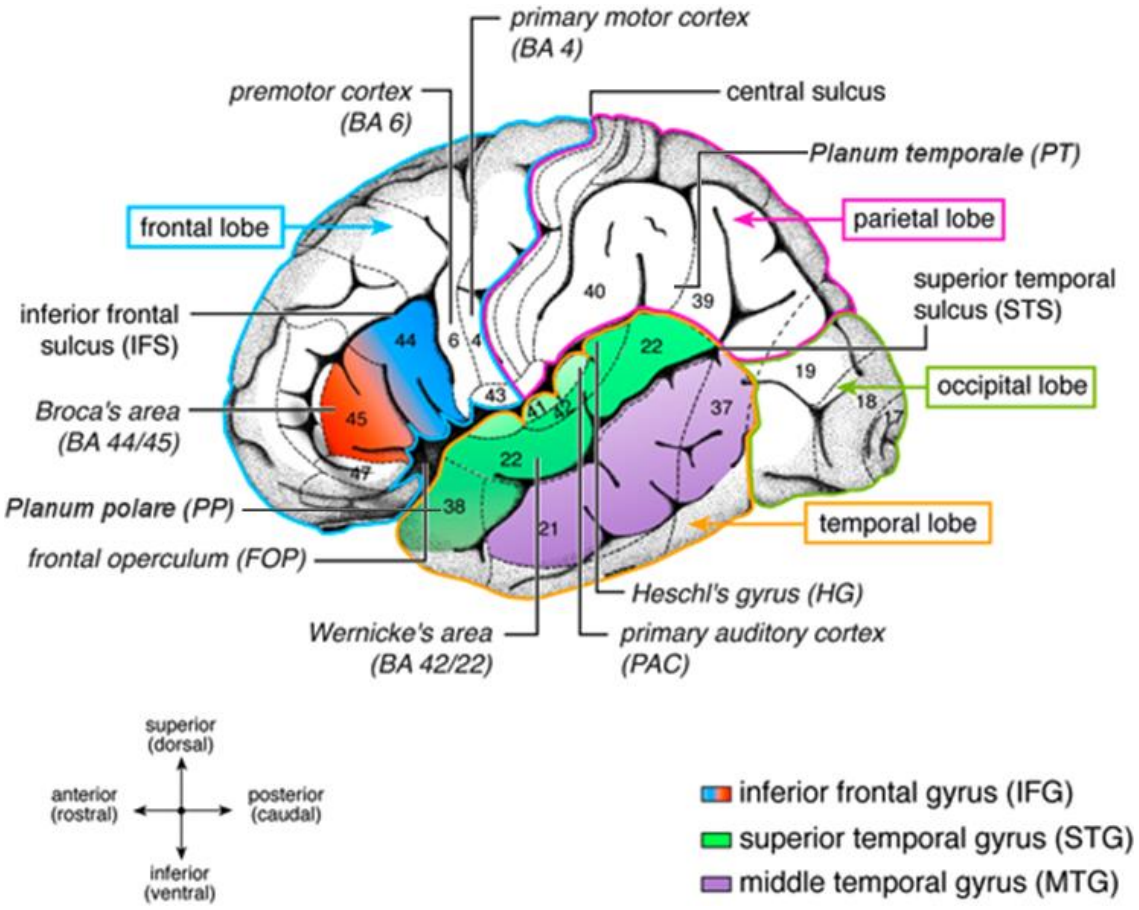


Figure 2.3: Anatomical details of the auditory processing network in the human cortex.
Original illustration from Friederici (2011). The different lobes (frontal, temporal, parietal, occipital) are marked by coloured borders. Major speech language relevant gyri (IFG, STG, MTG) are colour coded. Numbers indicate language relevant *Brodmann Areas* (BA; Brodmann, 1909). Broca's area consists of the *pars opercularis* (BA 44) and the *pars triangularis* (BA 45). The *premotor cortex* is located in BA 6. *Wernicke's area* is defined as BA 42 and BA 22. The *primary auditory cortex* (PAC) and *Heschl's gyrus* (HG) are located in a lateral to medial orientation.

Modern neuroimaging technologies such as magnetic resonance imaging (MRI), fMRI, EEG, MEG, and ECoG have been shown to be valuable tools for the investigation of speech processing. Recent work with these methods provided further insights into this phenomenon and revealed speech perception as a complex process that involves both left and right hemispheres and that is not limited to the temporal areas of the cortex (**Figure 2.3**) (Hagoort, 2005; Shalom and Poeppel, 2008; Friederici, 2011; Friederici, 2012; Hickok and Small, 2015). Furthermore, the interaction between speech comprehension and speech production, which involves also parietal regions of the cortex, was shown to be important and difficult to disentangle (Giraud et al., 2007).

A growing body of literature supports the hypothesis that speech perception is underpinned by a hierarchical system which processes features of speech at progressively higher levels of abstraction, from purely acoustic features of the speech sound to categorical units of speech (phonemes and syllables), words, and semantics (Hickok and Poeppel, 2004; Okada et al., 2010). However, there remains considerable controversy as to whether this analytic process proceeds strictly "left-to-right" in a feedforward manner, from smaller units (e.g., phonemes) to larger units (e.g., syllables; Chait et al., 2015). A growing body of research indicates that the human cortex is characterised by a massive amount of anatomical feedback versus feedforward connections (Gazzaniga, 2009). Although this seems to suggest a crucial role of top-down modulation in speech processing, the precise role of feedback is still hotly debated and will be discussed more fully in the next section.

Among the several models proposed to explain the cortical organisation of speech processing (e.g., Hickok and Poeppel, 2007; see **Figure 2.4**), some agreement is found in the hypothesis that there exists specialisation at the sound-word interface. In this sense, *lower-level* areas are the ones involved in the processing of the raw sound signal and the extraction of acoustical patterns from a speech sound, while *higher-level* areas are specialised in the processing of speech at the phonological, morphological, syntactical and lexical stages. One prominent model of the speech comprehension network was suggested by Hickok and Poeppel (2007), which macroscopically localised spectro-temporal analysis in dorsal STG and phonological level analysis in mid-posterior STS. Subsequently, this model suggests a divergence of this network into a *ventral stream*, which is involved in extracting and processing lexical components of speech and localises in MTG and ITS, and a *dorsal stream*, which maps speech onto articulatory motor representations (area Sylvian parietal temporal, between parietal and temporal lobes) and to more abstract and linguistic decision processes (IFG).

The cortical processing of speech has historically been attributed to the left or 'dominant' hemisphere. Although this is still hotly debated, recent research shed some light on this issue by supporting the conjecture that temporal processing at different timescales is associated with hemispherically asymmetric activation (Boemio et al., 2005; Jamison et al., 2006; Giraud et al., 2007; Obleser et al., 2008; Telkemeyer et al., 2009; Morillon et al., 2010; Giraud and Poeppel, 2012). One view proposes that right and left auditory areas are primarily suited to processing spectral and temporal changes respectively (Zatorre et al., 1992; Zatorre and Belin, 2001; Zatorre et al., 2002; Poeppel,

2003; Schonwiesner et al., 2005). Higher order areas such as STS have been suggested to exhibit a right hemispheric bias during the processing of sounds at the syllabic rate (200-300 ms; Boemio et al., 2005), whereas faster rates (25-50 ms) are thought to induce left lateralised responses (Poeppel, 2003; Zatorre and Salimpoor, 2013). In the context of speech, the emerging consensus is that cortical processing becomes progressively more left lateralised as signals become more speech-like, suggesting that laterality effects are more driven by higher order linguistic processing demands than by speech analysis *per se* (Binder et al., 2000; McGettigan et al., 2012; Peelle, 2012; Cogan et al., 2014; Overath et al., 2015).

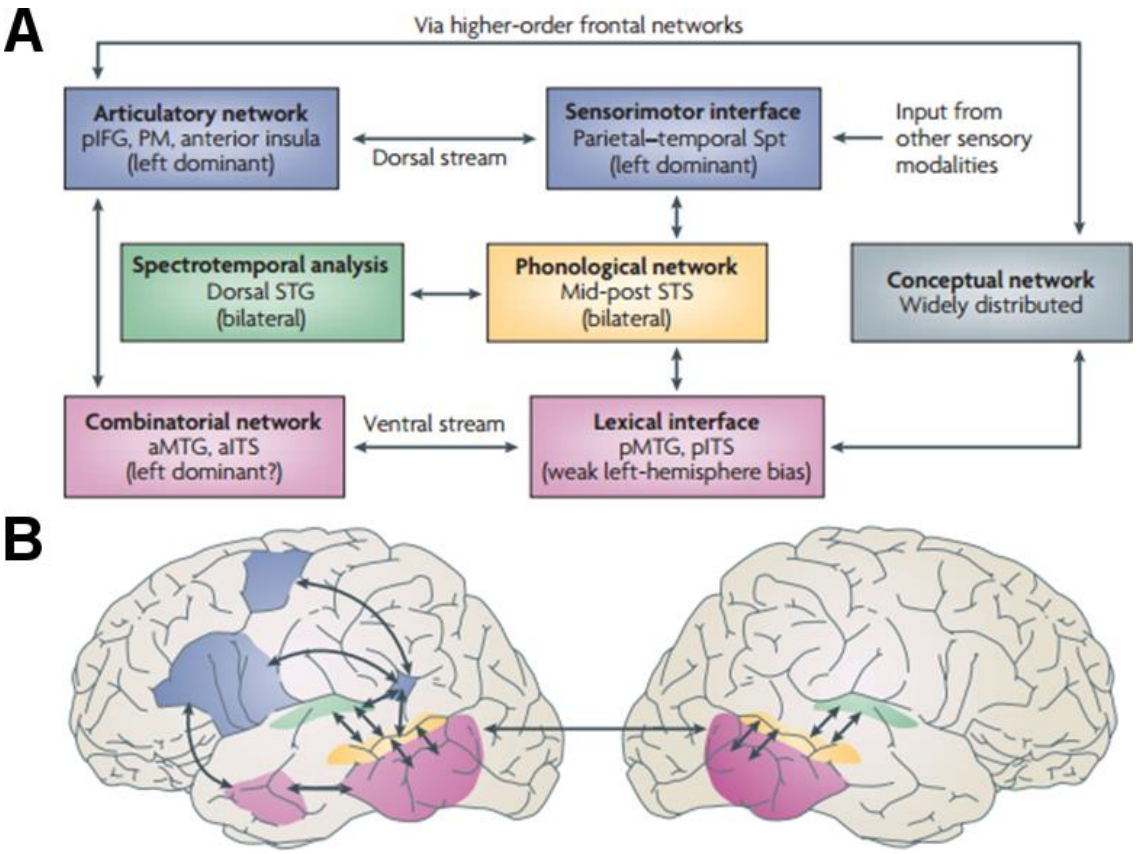


Figure 2.4: The dual-stream model of the functional anatomy of language.
Original figure from Hickok and Poeppel (2007). The schematic diagram (**A**) describes the main functional blocks of the dual-stream model, which are associated to specific areas of the cortex (**B**). The earliest stage of cortical speech processing involves some form of spectro-temporal analysis, which is carried out in auditory cortices bilaterally in the supra-temporal plane. These spectro-temporal computations appear to differ between the two hemispheres. Phonological-level processing and representation involves the middle to posterior portions of the STS bilaterally. Subsequently, the system diverges into a dorsal pathway (blue) that maps sensory or phonological representations onto articulatory motor representations, and a ventral pathway (pink) that maps sensory or phonological representations onto lexical conceptual representations.

2.1.3 Phoneme-level processing in human auditory cortex

Speech perception is the process that allows vocal communication and it consists of the transformation of a speech sound into its underlying meaning. An essential characteristic of this process is its robustness. In fact, it remains effective despite considerable natural variability across speakers and distortions in noisy and reverberant environments. Consequently, there is reason to believe that linguistic analysis cannot proceed directly from the sound waveform itself, but requires a processing stage that maps the variable speech sounds into invariant representations of speech units, such as phonemes and syllables (Liberman et al., 1967; Mesgarani et al., 2008; Overath et al., 2015).

In the field of linguistics, speech can be represented as a set of categorical phonological features that describe how a sound is articulated by humans (Chomsky and Halle, 1968; Jakobson et al., 1969). These features include properties of consonants, such as place of articulation (i.e., the location in the mouth where the constriction and obstruction of air occurs), manner of articulation (i.e., configuration and interaction of speech organs, such as tongue, lips, and palate), and glottal state of sounds (e.g., vibration, aspiration), as well as properties of vowels (e.g., back, high, and rounded vowels). Such properties can provide a "universal" feature set, i.e., language independent. Another way to represent speech is through a language-specific set of phonemes. Importantly, each phoneme maps to a unique configuration of phonetic features (**Table 2.1**).

Studies on small mammals showed that primary auditory areas encode spectro-temporal patterns that can discriminate features of speech such as the manner of articulation and voicing (Mesgarani et al., 2008), suggesting that humans and animals may build upon similar general acoustical structures to recognise and discriminate sounds. Crucially, there is evidence of the categorical encoding of such phonetic features in the STG, which may constitute a crucial step in the processing of a stimulus as speech rather than any other sound (Mesgarani et al., 2014; Leonard et al., 2016). The identification of categorical phonetic features allows for phoneme detection and, although it is unclear whether phonemes are explicitly encoded in the human cortex or by means of its constituent features, the processing leading to their categorical identification engages an extended cortical network that reaches from the auditory cortex to frontal and parietal regions (Scott and Johnsrude, 2003; Liebenthal et al., 2005; Turkeltaub and Coslett, 2010).

As expected by a hierarchical system, higher-level cortical areas in humans have been shown to be progressively more sensitive to more abstract features of speech and less susceptible to acoustic variation (DeWitt and Rauschecker, 2012; Overath et al., 2015). In particular, the region of STS (especially posterior regions) has been shown to be more sensitive to higher order intelligibility features and less to acoustic variations compared to auditory core regions (Okada et al., 2010). Also, STS has been suggested to be a possible area of speech-specific processing and encoding of linguistic features ranging from single phonemes to pairs of syllables (Hickok and Poeppel, 2007; Overath et al., 2015). However, there remains considerable uncertainty about its specific roles in the categorical perception and its integration in the speech comprehension network.

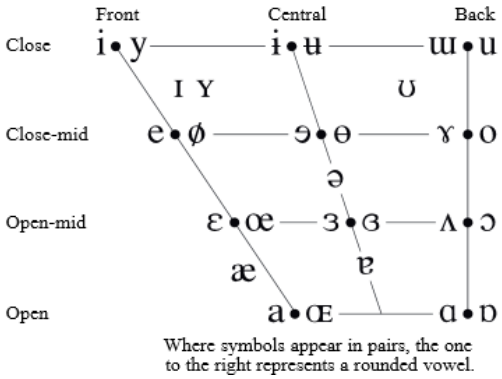
THE INTERNATIONAL PHONETIC ALPHABET (revised to 2015)

CONSONANTS (PULMONIC) © 2015 IPA

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k g	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			

Symbols to the right in a cell are voiced, to the left are voiceless. Shaded areas denote articulations judged impossible.

VOWELS



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

Table 2.1: Table of phonemes and phonetic features.
Adapted from *http://www.internationalphoneticassociation.org/content/ipa-chart*.

2.2 The integration of prior and cross-modal information in speech processing

Conversational speech is characterised by fast dynamics and large spectro-temporal variability. Humans can rapidly adapt to different voices and accents, and developed the remarkable ability to successfully perform speech comprehension despite the challenges of real world environments, such as noise, speech degradation, and competing speakers. In this context, there is widespread agreement that the perception does not only depend on the processing of sensory input, but that it benefits from the integration of information from different modalities and from prior knowledge. This section provides some insights on the current theories that explain such mechanisms of integration.

2.2.1 Principal factors contributing to speech perception

The previous section described the auditory system and presented current views on its functional organisation. In particular, the human cortex has been suggested to be organised in hierarchical stages that identify and process, from lower- to higher-levels, progressively more abstract elements of speech (Hickok and Poeppel, 2004; Okada et al., 2010; Peelle, 2012) (Section 2.1.2). In the simplest scenario, such a system could be represented as a unidirectional hierarchical network with a bottom-up information flow reflecting progressively more refined representation of speech extracted purely from sensory information (**Figure 2.5A**). This serial process would elicit cortical responses with precise temporal order, thus facilitating the isolation of neural indices at specific hierarchical levels along the time dimension. Although this pattern of cortical activation can be observed in response to simple isolated speech sounds (Salmelin, 2007), this basic model cannot account for several fundamental mechanisms required for natural speech comprehension, and thus it is not a good representation of our auditory system. For instance, lexically ambiguous words have multiple meanings, and the selection of the intended one may require prior knowledge or even information that has yet to be heard (Mason and Just, 2007). In the latter case, the listener (or reader) would have to wait for the following part of the speech to resolve such ambiguity, as is the case with the word *ball* in the following sentence:

*This time the **ball** was moved outside*
because the children had almost broken the window.

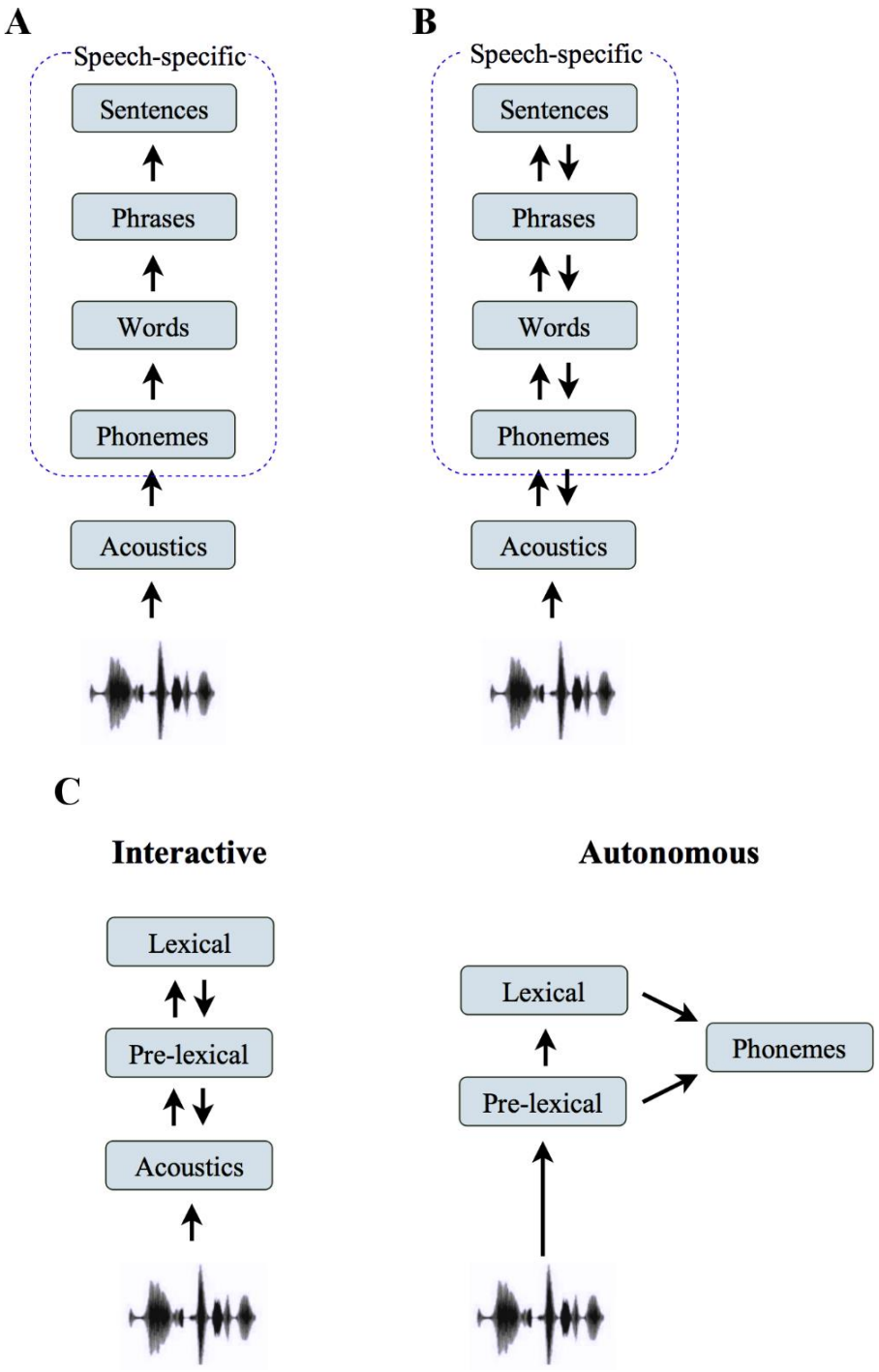


Figure 2.5: Schematic representations of information flow in models of speech perception.
(A) Feed-forward model of speech perception. This network describes the auditory system as fully characterised by the bottom-up processing of sensory information. **(B)** A feed-forward/feed-back model that enables the integration of prior-knowledge with sensory input. **(C)** This panel compares interactive and autonomous model rationales (adapted from McClelland et al., 2006). An interactive model posits bi-directional excitatory connections between processing levels with phoneme-level responses produced at the pre-lexical processing level (this schematic is a different view of the same model in panel B). An autonomous model of the sort advocated in Norris et al. (2000) posits strictly feedforward excitatory connections from pre-lexical to lexical processing and a separate stage of phoneme identification that combines information from both pre-lexical and lexical levels.

Similarly, the influence of semantic context on phonemic perception was studied in the context of phonemic ambiguity (Ganong effect, Ganong, 1980) and also when selected phonemes were replaced by a cough or noise (phonemic restoration effect, Warren, 1970). Studies based on these effects found that perception of the ambiguous segments is biased towards word sounds (rather than non-word ones) and toward words consistent with the prior semantic context (Connine and Clifton, 1987; Pitt and Samuel, 1993). These are examples of why the acoustic-phonetic encoding cannot be understood solely in terms of the mapping between an acoustic (or gestural) event and a phonetic percept. In this context, the most influential models of speech processing vary in the way that they incorporate sensory input with other sources of information, which is a crucial ability given that speech is usually heard under less-than-optimal listening conditions.

Speech comprehension benefits from various linguistic and acoustic properties that involve some types of past experience, for example about speaker, specific language, accent, listening conditions, or the semantic context. Past experience may come in the form of accurate prior knowledge or expectation, which provide sudden and dramatic changes in subjective perception (Ludmer et al., 2011; Sohoglu and Davis, 2016). For instance, strong perceptual enhancement of heavily distorted speech occurs when the listener has prior knowledge of the content, either exact words or semantic context (perceptual pop-out effect). Alternatively, past experience could entail a more gradual and incremental perceptual learning of regularities of the language, acoustic noise, or degradation over a timescale of minutes or longer (Sohoglu and Davis, 2016). For example, perception of speech that was degraded using noise-vocoding was shown to improve with experience (Goldstone, 1998; Ahissar and Hochstein, 2004; Davis et al., 2005). Although both perceptual learning and perceptual pop-out are based on prior experience, their very different effects may suggest that they originate from different brain mechanisms (Rubin et al., 1997; Norris et al., 2003). In particular, perceptual learning was linked to offline synaptic changes in low-level sensory processing, while prior knowledge and expectation were suggested to influence perception through activity at hierarchically late stages of processing.

One view of speech comprehension proposed autonomous models of speech that attribute most lexical effects to post-perceptual decisions and do not require feedback connections (Massaro, 1989; Norris et al., 2000). One model in particular, the *Merge model*, suggests the processing of phonemic codes and lexical analysis through two separate streams (Norris et al., 2000). In this model, information processed in pre-lexical

levels flows in a strictly bottom-up fashion to the lexical level, allowing activation of compatible lexical candidates. Crucially, information from both pre-lexical and lexical levels is available and merged at a subsequent decision stage, allowing explicit phoneme recognition with no need for perceptual effects between distinct levels. A different view suggests that a single system of integration including multiple interacting levels of processing supports the effects of both prior knowledge and perceptual learning (McClelland and Elman, 1986; Rubin et al., 1997; Friston, 2005; Mirman et al., 2006) (**Figure 2.5B**). This interactive processing view suggests that abstract higher-levels of the hierarchy (lexical levels) use prior-knowledge to facilitate speech processing at lower-levels (pre-lexical levels). The same interactions would lead to a synaptic change that facilitates a more effective processing of new presentations of similar stimuli. Several findings support this interactive approach and, specifically, the view that lexical factors can affect pre-lexical processing (see **Figure 2.5C** for a comparison between interactive and autonomous models; for a full review, see McClelland et al., 2006). However, the precise top-down/bottom-up mechanisms of integration within this interactive approach remain unclear (Norris et al., 2016).

2.2.2 Sharpening vs Predictive coding theories

The integration of sensory input and prior experience is not exclusive of auditory processing and has been extensively investigated in many sensory modalities (Kastner et al., 1999; Bar et al., 2006; Summerfield et al., 2006; van Ede et al., 2010; Zelano et al., 2011; Giraud and Poeppel, 2012; Kok et al., 2012a; Chennu et al., 2013; Büchel et al., 2014; Gardner and Fontanini, 2014). Previous research suggested that a single general neural mechanism may underpin this phenomenon. A growing body of literature has investigated a theory based on hierarchical Bayesian inference, in which higher order cortical regions modulate lower levels via top-down connections (Section 2.2.1), thereby facilitating sensory processing (Lee and Mumford, 2003; Friston, 2005; Yuille and Kersten, 2006; Summerfield and Koechlin, 2008). Within this framework, the critical issue is what does bottom-up information represent and what exactly is affecting prior experience through top-down connections.

One view is that higher hierarchical levels may suppress lower order sensory responses that are inconsistent with current expectation, which would result in a "sharpening" of the cortical sensory representation (Lee and Mumford, 2003). This could

be explained as being the result of feedback connections that either inhibit unexpected or enhance expected sensory information (Spratling, 2008). A different account of the Bayesian inference theory, named predictive coding, posits that higher order regions suppress the predictable information in lower processing areas and that, therefore, bottom-up connections propagate the prediction-error at each hierarchical stage (Mumford, 1992; Rao and Ballard, 1999; Murray et al., 2002). Also, a more recent model of predictive coding proposed the coexistence of prediction errors and sensory representation through two functionally distinct subpopulations of neurons: *Prediction error units* are associated with the signals that are passed forward from one level of the hierarchy to the next (i.e., the bottom-up signals); *Representation units* encode the putative sensory information, which can be sharpened through lateral interactions that subtract out the activity of the error units (Rao and Ballard, 1999; Friston, 2005; Jehee and Ballard, 2009; Clark, 2013). Thus, the updated information encoded in these representation units is used for the top-down predictions, which propagates to lower levels and generates prediction errors. By this account, this "delicate dance" between top-down and bottom-up appear to dissolve the superficially clean distinction between believing (i.e., prior knowledge, expectation) and perception, which emerge as being deeply mechanically intertwined (Clark, 2013).

One reason for excitement is the broad impact that these theories could have on the field of neuroscience. A better understanding of the architecture and functional organisation of cortical systems would lead to implications ranging from the interpretation of specific cortical responses (e.g., mismatch negativity responses, repetition suppression) to a better understanding of specific neural diseases (e.g., schizophrenia). In the specific case of "repetition suppression", multiple studies have shown a reduction in stimulus-evoked neural activity with stimulus repetition (Grill-Spector et al., 2006; Summerfield et al., 2008; Todorovic et al., 2011). In the context of predictive coding, this phenomenon could be addressed as a decrease of prediction error signals with the increase in predictability due to repetition. Please refer to Grill-Spector et al., 2006 for further discussion on the repetition suppression phenomenon, and to the work from Hohwy and Clark for additional discussion and interpretation of the predictive coding framework, also in the context of symptoms such as delusions and hallucination in schizophrenia (Clark, 2013; Hohwy, 2013; Clark, 2016).

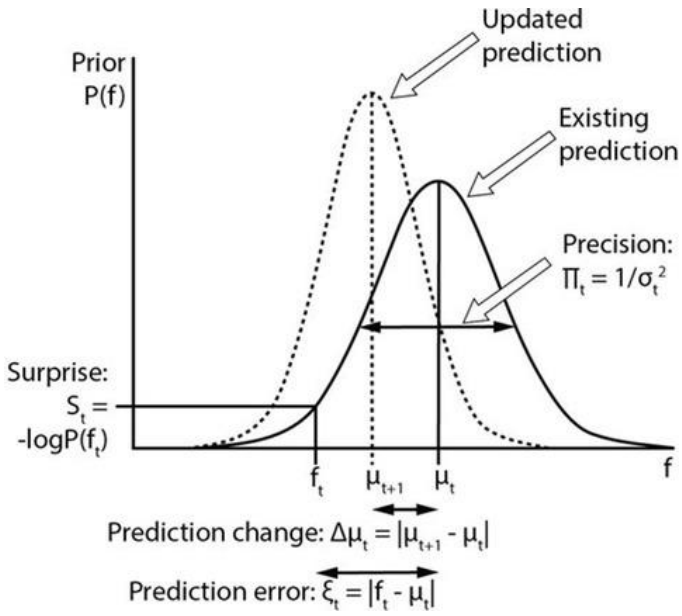


Figure 2.6: Computational variables involved in perceptual inference.
Original figure from Sedley et al. (2016). The graph represents a possible mechanism of perceptual inference in the context of auditory perception that is characterised by an active integration of expectation and prediction errors. This particular model refers to an auditory fundamental frequency detection task. The solid curve represents a schematic probability distribution of the prior prediction about the fundamental frequency (*f*) of an upcoming auditory stimulus (f_t), where *t* simply refers to the number or position of the stimulus within a sequence. This prediction is characterised by its mean (μ_t) and precision (Π_t), which is the inverse of its variance (σ_t^2). An unexpected f_t would result in a large prediction error (ξ_t), which is the absolute difference between μ_t and f_t. This mismatch between top-down prediction and bottom up sensory information results in a prediction update ($\Delta\mu_t$; i.e., Bayesian belief updating) that modifies the probability distribution of the expected sensory input (dashed line).

In recent years there has been some empirical support for both the sharpening and predictive coding accounts (Murray et al., 2002; Summerfield et al., 2008; den Ouden et al., 2009; Alink et al., 2010; Todorovic et al., 2011; Kok et al., 2012a; Kok et al., 2012b; Meyer et al., 2014), however there remains considerable controversy between these two models. In particular, one issue is that a reduction of neural activity in early sensory regions could agree with both the sharpening and predictive coding views, in fact that could be interpreted both as a suppression of expected or unexpected sensory information (Hsieh et al., 2010; Kok et al., 2012a). Also for this reason, most of the current evidence supporting either model tends to be indirect and focused on very specific domains, such as the testing of the Bayesian inference mechanism in specific modalities (**Figure 2.6**) (Ernst and Banks, 2002; Knill and Pouget, 2004). More recent studies have investigated this issue in a more direct way by considering indices of sensory representation, rather than just measures of the increase and decrease of cortical activation. For speech, it was demonstrated that an overall decrease of cortical activation may co-occur with an increase in sensory representation (**Figure 2.7**) (Murray et al., 2002; Egner et al., 2010; Blank and Davis, 2016; Sohoglu and Davis, 2016; Tuennerhoff and Noppeney, 2016). Although

these findings point to a predictive coding account rather than the sharpening hypothesis, more research is needed to clarify this issue in the context of auditory processing and, importantly, in the more general interpretation of the "predictive brains".

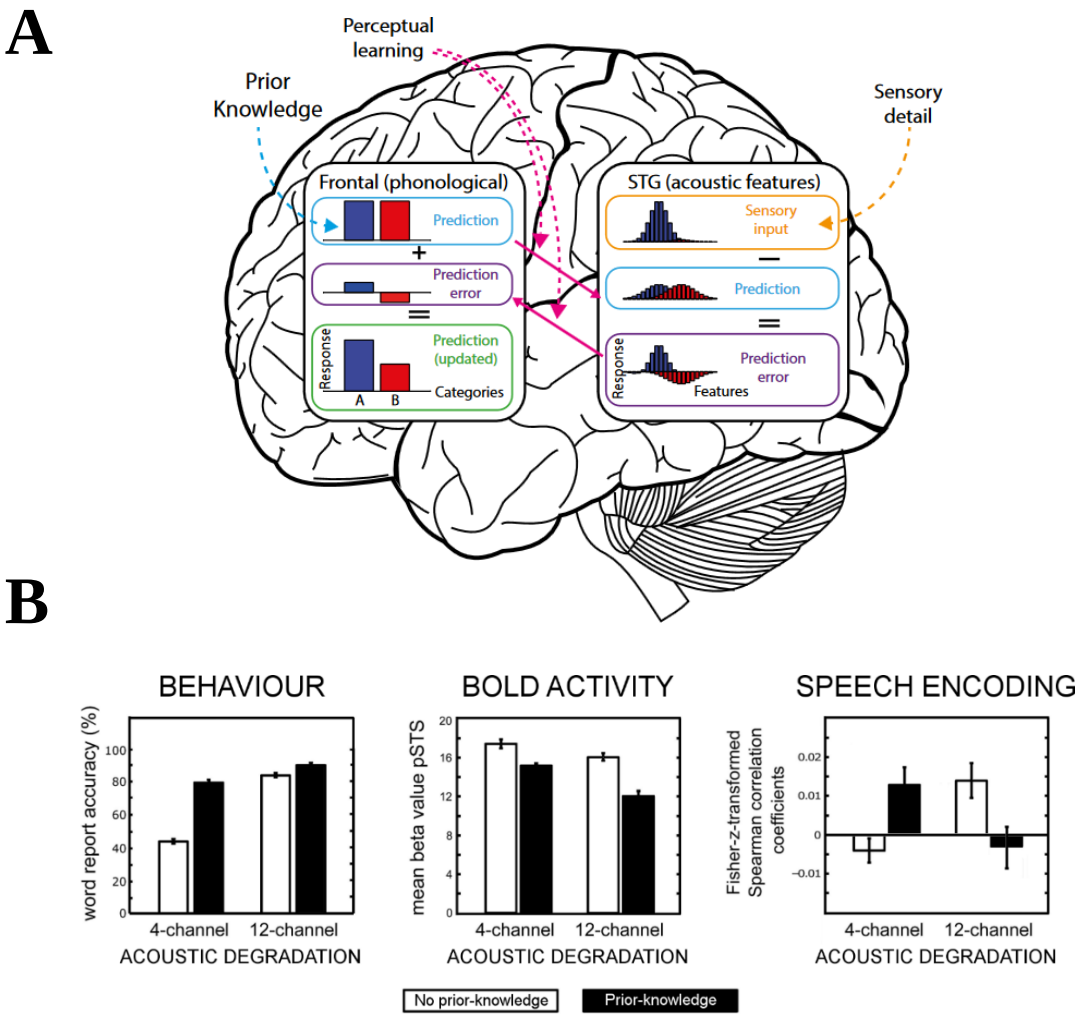


Figure 2.7: Predictive coding account for the speech processing network.
(A) The graph depicts a model of the speech comprehension cortical network in the context of phonological perception (Sohoglu and Davis, 2016). Behavioural and neural outcomes were determined by the interaction between two hierarchically organised levels of representation: sensory (acoustic-phonetic) features (STG) and phonological categories (frontal and somatomotor regions). The bar graphs represents perceptual hypotheses about the phonological content of speech (e.g., two phonological categories, A and B, colour-coded blue and red, respectively). Such predictions are conveyed top-down to the STG as predictions for upcoming sensory features via weights (solid pink arrows). The predictions are then updated using the previous prediction and the corresponding prediction error. In this scheme, perceptual learning (broken pink arrow) modulates the relative likelihood of the two categories encoded by activity in phonological prediction units. For each new sensory input, perceptual learning modulates the connection weights that map between phonological categories and acoustic features. **(B)** In this experiment, participants heard isolated words that were acoustically degraded using noise-vocoding. A higher number of frequency channels reflects a higher quality and, thus, better intelligibility. Prior knowledge on the upcoming stimulus (provided in text form) increased the perceived intelligibility for both 4- and 12-channel acoustic degradation. This perceptual increase corresponded to a decrease in BOLD activity in STS. In addition, an index of speech encoding reported a significant amount of speech information in STS only when speech was perceived as intelligible. Crucially, this was not the case for the 12-channel condition with prior knowledge, in which the prior information suppressed the speech encoding compared to when prior knowledge was not available (Adapted from Blank and Davis, 2016).

2.3 Measuring the cortical dynamics of speech perception

As described by the previous sections, speech comprehension involves a broadly distributed cortical network that exhibits both bottom-up and top-down dynamics. Although important insights have been recently provided about the functional roles of specific topographical areas of the cortex, this system is characterised by complex spatio-temporal dynamics and cross-frequency interactions that an approach based solely on spatial parcellation cannot capture. Because of this multifaceted nature of the speech processing network, several different technologies have been used to investigate distinct aspects of its underpinnings.

While much of our understanding of the neurobiology of language has been deduced from behavioural (Davis and Johnsrude, 2007) and lesion (Wernicke, 1874) studies, over the last 30 years or so those findings have been complemented by research using brain imaging. Firstly, primate invasive electrophysiology (Rauschecker and Scott, 2009) enabled the design of detailed models of the sound processing system. However, primates do not speak and the differences between their cortex and that of the human constitutes another obstacle to this type of imaging. Invasive brain recordings, such as ECoG, have also been performed on human subjects, providing neural data with good temporal and spatial resolution (Buzsáki et al., 2012; Mesgarani et al., 2014; Holdgraf et al., 2016; Leonard et al., 2016). However, it is usually performed on patients with severe cases of epilepsy just before they undergo brain surgery, which means that the broadly distributed speech processing network may be affected by such a condition. Furthermore, ECoG only samples a restricted area of the cortex (which is specific to each patient), meaning that the measurements cannot capture the network in its entirety. Another recent invasive approach that allows for the investigation of the functional segregation of cortical regions is focal cooling (Flinker and Knight, 2016; Long et al., 2016), which aims at altering speaking behaviour by cooling distinct cortical sites after craniotomy. However, this approach presents similar limitations as ECoG. fMRI overcomes these issues, as it does not have such strong limitations in terms of participant groups, and it allows for the imaging of the whole cortex (Logothetis, 2008; Glover, 2011; Overath et al., 2015; Blank and Davis, 2016; Tuennerhoff and Noppeney, 2016). However, fMRI has a limited temporal resolution that hampers its ability to investigate the fast dynamics of speech. Similarly, EEG and MEG enable a broad coverage that extends over the entire surface of the scalp, with no particular restrictions on the participant cohorts (Buzsáki et

al., 2012; Ding and Simon, 2014; Jackson and Bolger, 2014; Sohoglu and Davis, 2016; Baillet, 2017). Crucially, EEG and MEG provide very high temporal resolution at the cost of spatial resolution. Differently from intracranial recordings that can be pushed over a sampling rate of 10 kHz when aiming at the recording of neuronal spikes, non-invasive electro- and magneto-encephalography are usually conducted using a sampling rate of the order of 1 kHz. In fact, these technologies are strongly affected by a suppression at the high frequencies that follows the $1/f^i$ power law relationship (Keshner, 1982; Wornell, 1993). Finally, EEG and MEG record "direct" measures of electrical cortical activation, whereas fMRI measures variations of blood oxygen. For these reasons, EEG and MEG have the potential to provide new insights on the neurobiology of speech perception and are further discussed in this section.

2.3.1 Non-invasive electroencephalography

The brain is characterised by billions of interconnected neurons that form a complex network. The synchronous activation of neuron ensembles generates electric fields that can be measured from the scalp by placing electrodes on its surface. This procedure is completely non-invasive and the electrodes are usually in contact with the skin through a conductive gel. A number of electrodes are strategically placed on the scalp in order to cover most of the surface of the head, including the areas that are most representative of the neural activity of interest (**Figure 2.8**).

Due to the use of surface electrodes, the spatial resolution of an EEG system is limited to an inter-electrode distance of approximately 11 mm in high-density systems with 512 electrodes. Therefore, each electrode relates to the potentials generated by approximately 10^7 to 10^9 neurons and only macroscopic population effects can be measured. Furthermore, the layers of tissue between brain and scalp also cause the electrical response to smear across the scalp, an effect known as volume conduction (Freeman et al., 2003) which further reduces spatial precision.

A further level of complexity is caused by the small magnitude of brain signals, which is in the order of 20-40 μV. Therefore, the signal has to be amplified in order to be detected. Unfortunately, this operation is not selective and does also include unwanted signals such as electrooculogram (EOG) activity from eye blinking and movement (Corby and Kopell, 1972) and electromyogram (EMG) activity from muscle activation (Goncharova et al., 2003). We refer to these signals as EEG artifacts and the experiments

are carefully designed in order to reduce their occurrence during the recording and allow their identification and exclusion in the pre-processing stage which precedes the data analysis.

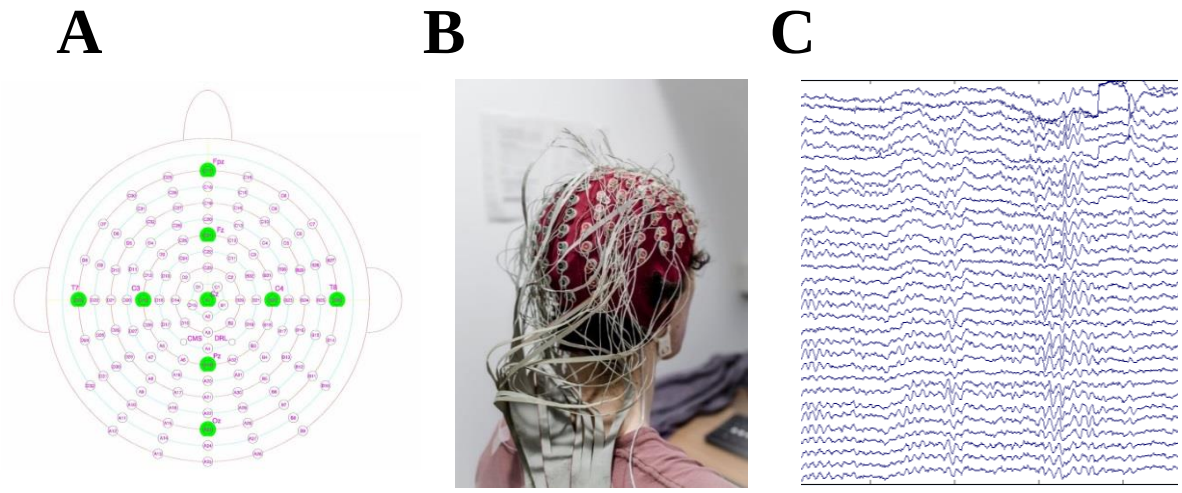


Figure 2.8: Electroencephalographic recording setup.
(A) Electroencephalographic measures are recorded using a number of scalp electrodes (usually 128 electrodes in this thesis; Courtesy of BioSemi B.V.). **(B)** These electrodes are in contact with the skin through a specific conductive gel and they are placed using an elastic cap according to a standard topographical distribution. **(C)** Voltage measures are recorded for each electrode and synchronised to the stimuli using specific trigger signals that indicates events such as the beginning of a sound file.

2.3.2 Magnetoencephalography

Electrical current produces magnetic induction, whose strength can be measured remote from the current source. The idea behind MEG is to measure the magnetic induction corresponding to the electrochemical current that flows within and between brain cells (Baillet, 2017). The magnetic signal produced by neural currents is on a scale of femtoteslas (10^{-15} T), over 10 million times smaller than the Earth's static magnetic field. For this reason, it is necessary for MEG technology to have high sensitivity and large dynamic ranges. Commercial systems feature coil magnetometers with whole-head coverage and about 300 independent channels, each sampled at up to 30 kHz. Superconductive temperatures are reached by enclosing the sensing apparatus in a thermally insulated tank (*dewar*) filled with liquid helium. Therefore, the MEG sensors are not in contact with the scalp, thus subject preparation times are much shorter than in EEG.

MEG technology presents a number of fundamental differences from EEG. Firstly, EEG signals are strongly affected by differences in electrical conductivity between the scalp, skull, and other biological tissues. This is not the case for MEG as magnetic

permittivity (the magnetic equivalent of conductivity) does not change for different layers. For this reason, MEG signals are less distorted than EEG electrical potentials and, in particular, allow for a clearer interpretation of the anatomical sources of the measured activity. However, the laws of physics impose that MEG signals decrease faster with source depth, which hampers the ability of this technology to capture activity from medial or subcortical brain regions. A second distinctive property of MEG is that it is sensitive mainly to quasi-tangential neural sources, while EEG can measure both quasi-tangential and quasi-radial sources (Cohen and Cuffin, 1987; Fuchs et al., 1998; Baillet et al., 1999; Ahlfors et al., 2010; Aydin et al., 2015). In addition, MEG signal depends on the location and orientation of the sensors with respect to neural sources. For this reason, both the monitoring of head movements within a session, and the registration of head positions between sessions and participants are crucial for data quality and group analysis and comparison. The extreme sensitivity of the MEG sensors means that they are affected by surrounding electromagnetic sources, such as metal objects or electrically powered instruments. Therefore, careful positioning of the device is important for signal quality. Finally, unlike EEG data, MEG measures are reference-free.

2.3.3 Event-Related framework

A primary use of electro- and magnetoencephalography is the study of brain responses to sensory stimuli. In these studies, the participant is usually presented with a number of stimuli and may be asked to perform a task such as answering questions or pressing a button in response to a particular target. The precise timing of these events has to be known (or recorded, if randomised) and synchronised with the neural recording. This allows the experimenter to associate a brain response to the stimulus that caused it and to study the temporal dynamics of the brain activity. Such brain responses are called event-related potentials (ERP) for EEG, and event-related fields (ERF) for MEG. However, the signal-to-noise ratio (SNR) of the recorded signal is usually insufficient to study responses to single events (see Sections 2.3.1 and 2.3.2). In fact, the signal of interest that reaches the EEG or MEG sensors has a very small magnitude and is mixed with other unwanted signals, such as spontaneous neural activity, muscular electrical activity due to movements of the participant, and external electrical noise.

One of the most common methods to overcome the SNR issue is to extract a series of EEG/MEG epochs which are time-locked to a repeated stimulus and to average them.

By assuming noise has a zero mean and that a particular stimulus elicits the same ERP/ERF at each presentation, this averaging technique can be used to reduce noise while preserving the response to the stimulus (Luck, 2005).

ERP/ERF components (peaks and troughs) are studied in terms of their characteristic amplitude, latency, and topographical distribution. In terms of auditory processing, the ERP/ERF elicited in response to sound stimuli is known as an auditory evoked potential/field (AEP/AEF; **Figure 2.9**). The latencies of such auditory responses can be grouped into three parts: early responses come from the brainstem (auditory brainstem response - ABR); the middle-latency responses are derived from an initial activation of the auditory cortex; late responses come from auditory and associated cortices (Picton, 2013). Notably, research on AEPs/AEFs in response to speech suggests that components with longer latencies relate to progressively higher levels of the speech processing hierarchy (Salmelin, 2007). This indicates that AEP/AEF are closely tied to perception as they reflect more than the low-level acoustics of the stimulus. For this reason, the use of auditory evoked responses as a clinical tool in audiology has been suggested in the context of auditory threshold estimation, indexing the auditory system development, auditory discrimination, and fine tuning of auditory implants such as cochlear implants (Cone-Wesson and Wunderlich, 2003; Lopez-Valdes et al., 2013; Paulraj et al., 2015).

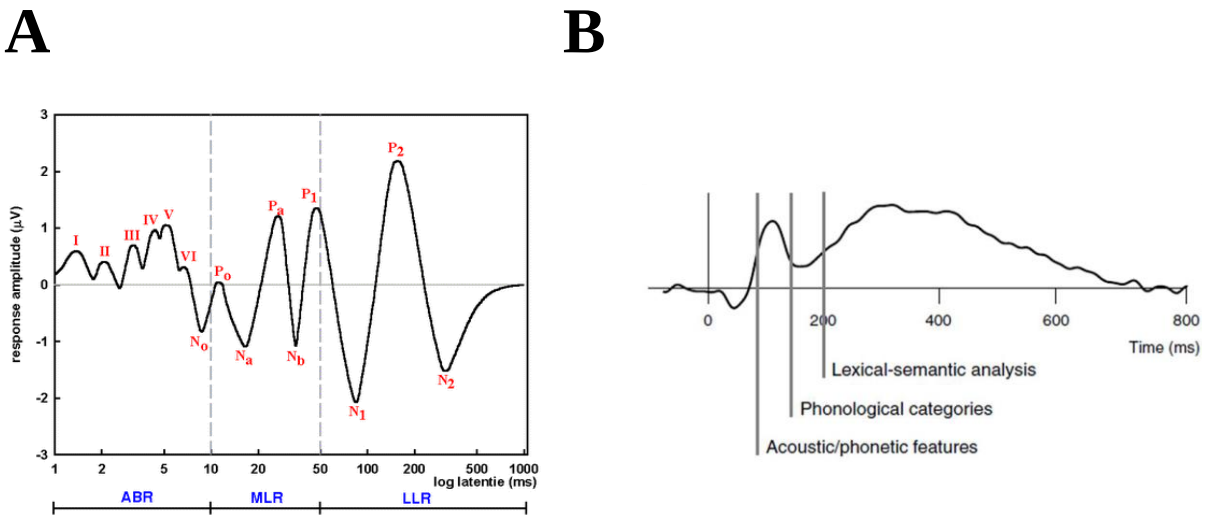


Figure 2.9: A canonical auditory evoked potential (AEP).
(A) Peaks I–VI represent the auditory brainstem response (ABRs), peaks N_0–N_b represent the middle latency auditory evoked potentials (MLR) and peaks P_1–N_2 represent the long latency response (LLR) (Picton et al., 1974). **(B)** Time course of auditory-evoked responses to speech in superior temporal cortex (Salmelin, 2007).

2.3.4 Modelling the response to continuous speech

In the last few decades, ERP/ERF techniques have yielded great insight into the processing of sounds and speech in the human brain. These methodologies are effective when short, isolated, and frequent stimuli are presented which, in the context of speech, can be syllables or words. However, the event-related framework is ill-suited for studying cortical responses to naturalistic stimuli such as continuous speech. In American English natural speech, units of sounds such as syllables occur at a rate of about 3 to 10 Hz, which correspond to durations from 100 to 330 ms (**Figure 2.10**). Importantly, there exist even smaller units of speech, called phonemes, which occur at faster rates that range from about 6 to 32 Hz, meaning that some units have durations as low as 30 ms (Kuwabara, 1996; Crosse et al., 2016a). Such fast dynamics of natural speech imply that long latency responses to a phoneme/syllable overlap with short latency activity elicited by the following unit. This overlap hampers the effectiveness of the event-related approach in the study of continuous speech.

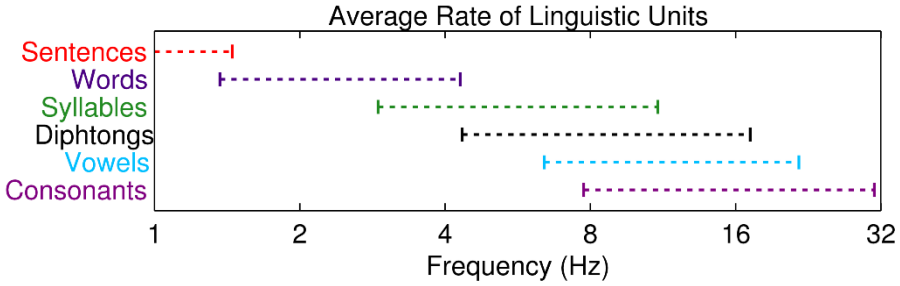


Figure 2.10: Average rate of linguistic units.
This result was based on the normal speaking rate of American English. The brackets indicate the mean ± SD (Kuwabara, 1996; Crosse et al., 2016a).

The human auditory system is tuned to process continuous speech. In fact, short isolated (discrete) sounds have been shown to elicit different patterns of cortical responses (Bonte et al., 2006). For this reason, recent studies have investigated approaches for analysing EEG/MEG responses to continuous sounds (Lalor et al., 2009; Lalor and Foxe, 2010; Ding and Simon, 2014). The focus has been on the time scale most critical for speech recognition, which is in the order of hundreds of milliseconds (1-10 Hz). One way of representing the temporal fluctuations at this time scale is the *amplitude envelope* (also known as the *temporal envelope* or, more simply, *envelope*), which summarises a speech or non-speech sound as a univariate signal in time. Research studies in animal neurophysiology has reported that signals from single neurons in primary auditory cortex encode the envelope of non-speech sounds by phase-locked neural firing (Wang et al.,

2003). This cortical entrainment phenomenon has been demonstrated for the envelope of speech in humans using EEG (Aiken and Picton, 2008), MEG (Ahissar et al., 2001; Luo and Poeppel, 2007), and ECoG (Nourski et al., 2009). In particular, the envelope entrainment has been shown to be prominent for low frequency neural activity (< 8 Hz; delta- and theta-band, 1-4 and 4-8 Hz respectively) and for the amplitude envelope of high-gamma band activity (Pasley et al., 2012; Zion Golumbic et al., 2013b).

Although the cortical entrainment to the envelope of speech is now well established, its underlying neural mechanisms and functional roles remain controversial (Ding and Simon, 2014). This phenomenon has been shown for non-speech sounds and for unintelligible speech (Lalor et al., 2009; Howard and Poeppel, 2010; Millman et al., 2013; Steinschneider et al., 2013), indicating contributions from lower-level areas related to the processing of the sound acoustics. Furthermore, recent studies have revealed the contribution of higher-level processing areas. In particular, top-down cognitive functions such as attention have been shown to modulate the cortical entrainment to speech (Kerlin et al., 2010; Ding and Simon, 2012a; Mesgarani and Chang, 2012; Zion Golumbic et al., 2013b; O'Sullivan et al., 2014). Importantly, it has been suggested that the entrainment to continuous speech depends also on the listeners' ability to extract linguistic information that, as described in Section 2.2, may involve more than just sensory information (Peelle et al., 2013; Zoefel and VanRullen, 2015).

A number of essential questions remain unanswered about the nature of the envelope entrainment phenomenon. Firstly, although there is agreement about the contribution of lower-level auditory areas, it remains unclear which acoustic features are driving the cortical responses (e.g., acoustic "edges", pitch, coarse spectro-temporal modulation). Furthermore, in terms of the contribution from the higher-level stages of the speech processing hierarchy, unresolved questions remain about what aspects of the top-down cognitive functions are reflected in the neural responses, for instance whether the main contribution comes from the modulation of activity in lower-level areas or also from activity in non-primary areas.

The anatomical and functional understanding of this phenomenon and the ability to isolate contributions from different hierarchical levels and/or cortical areas has the potential to provide significant advances in the study of speech processing in the human brain and it is a central topic of this thesis. To this end, it is crucial to identify analysis approaches that best exploit the specific type of neural recording used. Cortical entrainment can be quantified using measures of cross-correlation or coherence between

the speech envelope and the neural recordings (Peelle et al., 2013; Millman et al., 2015; Thwaites et al., 2015; Baltzell et al., 2016). An alternative and potentially more sensitive approach consists of estimating the mapping between the speech envelope and recorded neural signals, such as the spectro-temporal response function (STRF) method (Depireux et al., 2001; Ding and Simon, 2012a). The following subsection describes a recent methodology for estimating such a mapping, which allows for the investigation of the processing of continuous auditory stimuli.

2.3.5 Temporal response function (TRF)

The human brain is highly non-linear and time-variant by nature. However, a large number of studies have brought new insights to the state-of-the-art by assuming that it (or some parts of it) behaves like a linear time-invariant (LTI) system (Bialek et al., 1991; Stanley et al., 1999; Lalor and Foxe, 2010; Ding and Simon, 2012b; Mesgarani and Chang, 2012; Pasley et al., 2012). This assumption may be very strong, but it allows the use of modelling approaches such as linear regression. An advantage of this framework is to have a system entirely described by its impulse response. Such impulse responses can be used as dependent measures of speech processing and can be estimated, given the input and output signals, using a number of different techniques. Moreover, when the impulse response is available (the real one or its estimation), it can be used to predict the output given the input (*forward-modelling*) or to reconstruct the input given the output (*backward-modelling*). We refer to the estimated system response as a temporal response function (TRF) and to this framework as the TRF approach (Ding and Simon, 2012a; Crosse et al., 2016b).

If the system that we want to model is the human brain and, specifically, its subsystems that are involved in sound and speech processing, then the input is the auditory stimulus presented to the participant while the output is the neural signal recorded (e.g., EEG, MEG, ECoG). The forward model can be represented in discrete time as:

$$r(t,n) = \sum_{\tau} w(\tau,n)s(t-\tau) + \varepsilon(t,n), \qquad (2.1)$$

where $\varepsilon(t,n)$ is the residual response at each channel *n* and time point *t* not explained by the model. The TRF, $w(\tau,n)$, can be thought of as a filter that describes the linear transformation of the ongoing stimulus $s(t-\tau)$ to the ongoing neural response $r(t,n)$.

This transformation is described for a specified range of time lags τ relative to the instantaneous occurrence of the stimulus feature $s(t)$.

A number of recent studies have shown that regularised linear regression is an effective method for system estimation (Lalor et al., 2009; Power et al., 2011; O'Sullivan et al., 2014; Crosse et al., 2016a). In particular, this approach has been shown to be effective for studying the spatiotemporal dynamics of the non-invasively recorded cortical signal in response to speech stimuli. These dynamics can be studied using the regression weights returned by the forward model estimation. These weights, $w(\tau, n)$, are referred to as TRF and are estimated (as for a standard linear regression) by minimising the mean-square error (MSE) between the neural response $r(t, n)$ and its prediction $\hat{r}(t, n)$:

$$min\big(\varepsilon(t,n)\big) = \sum_{t}[r(t,n) - \hat{r}(t,n)]^2. \tag{2.2}$$

This minimisation problem can be solved by applying the following closed formula (de Boer and Kuyper, 1968):

$$w = (s^T s)^{-1} s^T r. \tag{2.3}$$

This formula is valid in the basic case of a single input and an instantaneous relationship between input and output. However, although the processing of a speech sound in the human brain is very fast, it is not instantaneous. Therefore, the TRF has to include a certain time-lag window of a few hundred milliseconds, starting from the presentation of a stimulus. A comparison between forward TRF and AEP responses can be obtained by matching this time-window with the epoch length of the AEP epochs. For example, in the context of sound perception, a window of interest could extend from -100 (τ_{min}) to 400 ms (τ_{max}), where the lag -100 ms would correspond to the cortical response to a stimulus 100 ms before that stimulus was presented, whereas the TRF at 400 ms would index how a change in the input sound would affect the EEG after 400 ms. In this context, the forward TRF function shows a high degree of correspondence with the common AEP in terms of its major components and their time course (Lalor et al., 2009). Unlike the event-related framework, this approach allows for the study of the responses to a continuous stimulus (Mesgarani et al., 2009) and, importantly, it allows the estimation of responses to two or more stimuli presented concurrently (Power et al., 2012).

The inclusion of a time-lag window is obtained simply by replacing the stimulus vector *s* in Equation 2.3 with a matrix *S* containing the time delayed versions of that same stimulus:

$$S = \begin{bmatrix} s(1-\tau_{min}) & (-\tau_{min}) & \dots & s(1) & 0 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots & s(1) & \dots & \vdots \\ \vdots & \vdots & \dots & \vdots & \vdots & \dots & 0 \\ \vdots & \vdots & \dots & \vdots & \vdots & \dots & s(1) \\ s(T) & \vdots & \dots & \vdots & \vdots & \dots & \vdots \\ 0 & s(T) & \dots & \vdots & \vdots & \dots & \vdots \\ \vdots & 0 & \dots & \vdots & \vdots & \dots & \vdots \\ \vdots & \vdots & \dots & \vdots & \vdots & \dots & \vdots \\ 0 & 0 & \dots & s(T) & s(T-1) & \dots & s(T-\tau_{max}) \end{bmatrix}. \quad (2.4)$$

The resulting formulation includes the inversion of the matrix $\mathbf{S^TS}$ (the autocovariance matrix), which is an operation particularly prone to numerical instability when solved with finite precision. This problem can be solved by introducing a bias term λ or 'smoothing factor' to the TRF estimation formula:

$$w = (S^TS + \lambda I)^{-1}S^Tr, \quad (2.5)$$

where **I** is the identity matrix. Addition of this smoothing term also solves the other main issue, that of overfitting. In fact, this additional weighting of the diagonal of $\mathbf{S^TS}$ before matrix inversion, a procedure known as Tikhonov regularisation (Tikhonov et al., 1977), prevents overfitting to high-frequency noise along the low-variance dimensions (Theunissen et al., 2001; Mesgarani et al., 2008).

Indeed, there are other approaches that allow the handling of continuous stimuli, such as cross-correlation and coherence analyses. However, these are better suited to stimuli modulated by a stochastic process such as Gaussian white noise, while natural stimuli such as continuous speech rarely conform to a white random process. For this reason, the TRF represents a more temporally precise method to characterise sensory systems in response to naturalistic stimuli such as speech (please refer to Lalor et al., 2009; Crosse et al., 2016b for experimental evidence). Furthermore, the TRF has the important advantage (over both cross-correlation and coherence analyses) of being a predictive model, which facilitates the acquisition of quantitative measurements based on prediction (of the neural data) and reconstruction (of the stimulus). In particular, if a neural recording is split into two parts, trials_A and trials_B, a TRF fit on trials_A could be used to predict the EEG signal or to reconstruct the input stimulus of trials_B. This approach, which has been discussed in the context of brain-computer interfaces

(O'Sullivan et al., 2014; Mirkovic et al., 2015), allows the quantitative estimate of the quality of the model derived. This is achieved using leave-one-out cross validation to generate predictions over the whole dataset, which are then compared with the actual data (EEG signals and input stimuli for forward and backward models respectively) (Crosse et al., 2016b). This produces a measure that reflects the quality of the estimated model that can be used to compare distinct conditions or subject groups.

Although linear regression approaches, such as the TRF, have been shown to effectively model the neural responses to natural speech, current and past research has been focusing on the cortical entrainment to the acoustic envelope of speech (Ding and Simon, 2014). The following chapter aims at clarifying whether non-invasive EEG and MEG reflect more than the passive following of the low-level speech acoustics and investigates specifically the cortical processing at the level of phonemes.

Chapter 3. Low frequency cortical entrainment to speech reflects phonemic level processing

3.1 Introduction

Human brains have evolved to convert the spectro-temporally complex acoustic patterns of natural speech into coherent, categorical, semantic representations. This ability is one of the characteristics that most strikingly distinguishes humans from other species. However, we are far from understanding the precise cortical mechanisms that allow humans to accomplish this task. As set out in Chapter 1, the overarching goal of this thesis is to define non-invasive and robust procedures to isolate neural indices of natural speech processing at the level of phonemes. EEG is the best candidate to achieve this objective, as it is a relatively inexpensive, non-invasive technology that provides us with direct recordings of cortical activity with high-temporal resolution. The latter factor especially makes EEG a suitable solution when using sensory stimuli with fast dynamics such as speech. As detailed in Section 2.3, EEG signals have been shown to reflect the cortical tracking of the envelope of speech, however it remains unclear whether this phenomenon simply reflects low-level responses to speech acoustics or if it actually includes processing of speech-specific features.

Recent research has provided important insights into this fundamental question by identifying specific areas of the cortex that are involved in phoneme-level processing. In particular, phonetic features have been shown to be encoded in the superior temporal gyrus (Mesgarani et al., 2014), which suggests a role of that area in the transformation of speech acoustics into categorical representations. Another temporal area thought to be involved in this processing stage is the superior temporal sulcus, whose function has been

shown to be distinct from lexical, semantic or syntactic processes (Overath et al., 2015). These findings, which are in line with the hypothesis of a hierarchical organisation of the speech processing network, suggest that categorical speech information is encoded in cortical areas that are potential generators of EEG signals (Huang et al., 2003; Schonwiesner et al., 2007; Ruhnau et al., 2013).

This chapter investigates the degree to which low-level (envelope, spectrogram) and higher-level (phonemic, phonetic feature) characteristics of natural speech are reflected in EEG activity. In doing so, we provide evidence that EEG does not just reflect passive neural following of the acoustic energy of speech, but that it also indexes the categorical perception of phonemes in the human brain. Furthermore, we sought to determine whether processing of different phonetic features can be discriminated in EEG responses and found that this discriminative power varies as a function of response latency, in line with what one might expect of a hierarchical system. The findings described in this chapter were presented at several international conferences and were published as: "Low frequency cortical entrainment to speech reflects phoneme level processing", *Current Biology*, 25:2457-2465, October 2015.

3.2 Methods

3.2.1 Subjects

This study consisted of two experiments. Ten healthy subjects (7 male) aged between 23 and 38 years old participated in the first experiment, and ten healthy subjects (7 male) aged between 21 and 32 years old participated in the second (5 subjects participated in both experiments). The study was undertaken in accordance with the Declaration of Helsinki and was approved by the Ethics Committee of the School of Psychology at Trinity College Dublin. Each subject provided written informed consent. Subjects reported no history of hearing impairment or neurological disorder.

3.2.2 Stimuli and Experimental Procedure

In the first experiment, subjects undertook 28 trials, each of the same length (just under 155 seconds), where they were presented with a professional audio-book version of a popular mid-20th century American work of fiction written in an economical and understated style and read by a single male American speaker. The trials preserved the

storyline, with neither repetitions nor discontinuities. The average speech rate was ~210 words/min. Similarly, the second experiment involved the presentation of the same trials in the same order, but with each of the 28 speech segments played in reverse. All stimuli were presented monophonically at a sampling rate of 44,100 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 centred on the screen, and to minimise eye blinking and all other motor activities.

3.2.3 Data Acquisition and Preprocessing

Electroencephalographic (EEG) data were recorded from 128 scalp electrodes for 18 subjects, and 160 scalp electrodes for 2 subjects. The data acquired using the 160-electrode system were mapped to the same 128 electrode positions used for all other subjects using a spline interpolation algorithm (EEGLAB; Delorme and Makeig, 2004) resulting in a coherent dataset with identical channel configuration for all subjects. Data were filtered over the range 0 - 134 Hz, and digitised with a sampling frequency of 512 Hz using a BioSemi Active Two system. Data were analysed offline using MATLAB 2014 software (The Mathworks Inc.). EEG data were digitally filtered between 1 and 15 Hz using a Chebyshev Type 2 filter in both the forward and backward directions to remove phase-distortion. In order to reduce the processing time required, all EEG data were then down-sampled to 128 Hz. Excessively noisy EEG channels were rejected based on several criteria (Junghofer et al., 2000), and the data on these channels were estimated using spherical spline interpolation (EEGLAB; Delorme and Makeig, 2004). Independent component analysis (ICA) was performed independently for each subject using the Infomax algorithm (Makeig et al., 1996). Components constituting artifacts were removed via visual inspection of their topographical distribution and frequency content and the remaining components were back-projected to EEG electrode space. All channels were then referenced to the average of the two mastoid channels.

3.2.4 TRF computation

The method used here to analyse the mapping between the various speech representations and the recorded EEG data is commonly known as a temporal response function (TRF; Section 2.3.5). A TRF can be interpreted as a filter that describes the brain's linear transformation of a stimulus feature, *s(t)*, to the continuous neural response *r(t)*, i.e.,

$$r(t) = TRF * s(t), \quad (3.1)$$

where * represents the convolution operator. The TRFs were calculated by performing regularised linear regression between our stimulus variables and our EEG. Specifically we perform ridge regression wherein a parameter (λ) is set to control overfitting (see Section 2.3.5 and Crosse et al., 2016b for a detailed description of this step). Given that the stimulus here is often represented as a multivariate feature vector, we refer to our TRFs as multivariate TRFs (mTRFs). mTRFs were calculated using custom written, publicly available software (*https://sourceforge.net/projects/aespa/*, Crosse et al., 2016b).

Previous work attempted to cast TRF functions with μV as their unit of measure (Lalor et al., 2006; Lalor et al., 2009). However, this relies on a decision to normalise the input stimulus values between some limits and, as such, has been somewhat arbitrary. In addition, in the present work, the mTRFs are multivariate which further complicates the issue of precise units. For these reasons, and in line with previous work from other groups (e.g. Ding and Simon, 2012a), the mTRFs are presented here in arbitrary units. The colours in the mTRF plots (**Figures 3.4** and **3.5**) can be interpreted as follows: red at a particular latency indicates that, at that post-stimulus lag, the EEG voltage is driven in a positive direction by the presentation of that particular stimulus (e.g., phoneme or frequency). And blue means the EEG voltage at that post-stimulus lag is driven negative by that stimulus. Thus, given the same normalisation strategy for forward and time-reversed speech, the mTRF responses can be compared in terms of their amplitudes, despite their description in terms of arbitrary units.

3.2.5 Speech Representations

Linear regression was used to determine multivariate temporal response functions (mTRFs) describing a mapping between the EEG and five speech representations:

1. Broadband amplitude envelope **(E)**: This was calculated as:

$$E = |x_a(t)|, \quad x_a(t) = x(t) + j\hat{x}(t), \tag{3.2}$$

where $x_a(t)$ is the complex analytic signal obtained by the sum of the original speech $x(t)$ and its Hilbert transform $\hat{x}(t)$. E was defined as the absolute value of $x_a(t)$. This was then downsampled to the same sampling frequency as the EEG data, after applying a zero-phase shift anti-aliasing filter.

2. Spectrogram **(S)**: This was obtained by first filtering the speech stimulus into 16 frequency bands between 250 Hz and 8 kHz according to Greenwood's equation (Greenwood, 1961), and then computing the amplitude envelope (as above) for each frequency band.

3. Phonemes **(P)**: This representation was computed using the *Prosodylab-Aligner* (Gorman et al., 2011) which, given a speech file and the corresponding textual orthographical transcription, automatically partitions each word into phonemes from the American English International Phonetic Alphabet (IPA) and performs forced-alignment (Yuan and Liberman, 2008), returning the starting and ending time-points for each phoneme. This information was then converted into a multivariate time-series composed of indicator variables, which are binary arrays (one for each phoneme). These are active for the time-points in which phonemes occurred. The phonemes are mutually exclusive, so that only one can be active at each sample point. We selected a subset of the IPA composed of the 35 most frequent phonemes in the presented speech stimuli (3 of 38 IPA phonemes were excluded as being outliers in terms of how rare they were). P is a language dependent representation of speech.

4. Phonetic features **(F):** This representation was obtained through a linear mapping of the phonemic representation into a space of 19 features (based on the University of Iowa's phonetics project *http://www.uiowa.edu/~acadtech/phonetics/english/english.html/*) and using an approach similar to Mesgarani et al. (2014). This set of phonetic features were a distinctive subset of those defined by Chomsky and Halle (Chomsky and Halle, 1968) to describe the articulatory and acoustic properties of the phonetic content of speech. In particular, the chosen features are related to the manner

of articulation, to the voicing of a consonant, to the backness of a vowel, and to the place of articulation. Each phoneme consists of a combination of distinct features; therefore this is a set of non-mutually exclusive descriptors. F is a language independent representation of speech.

5. Finally, we propose a model that combines F and S (**FS**): This was obtained by concatenating F and S into a single data matrix. This representation consists of 19 phonetic features and 16 frequency bands; therefore FS has 35 dimensions. The rationale for combining these particular two representations was that the better performance of the S-model relative to the E-model suggested it as a more optimal way to capture processing of the low-level acoustics. Choosing between P and F was simply done for efficiency given that F is essentially a more concise representation of the same information as that contained in P.

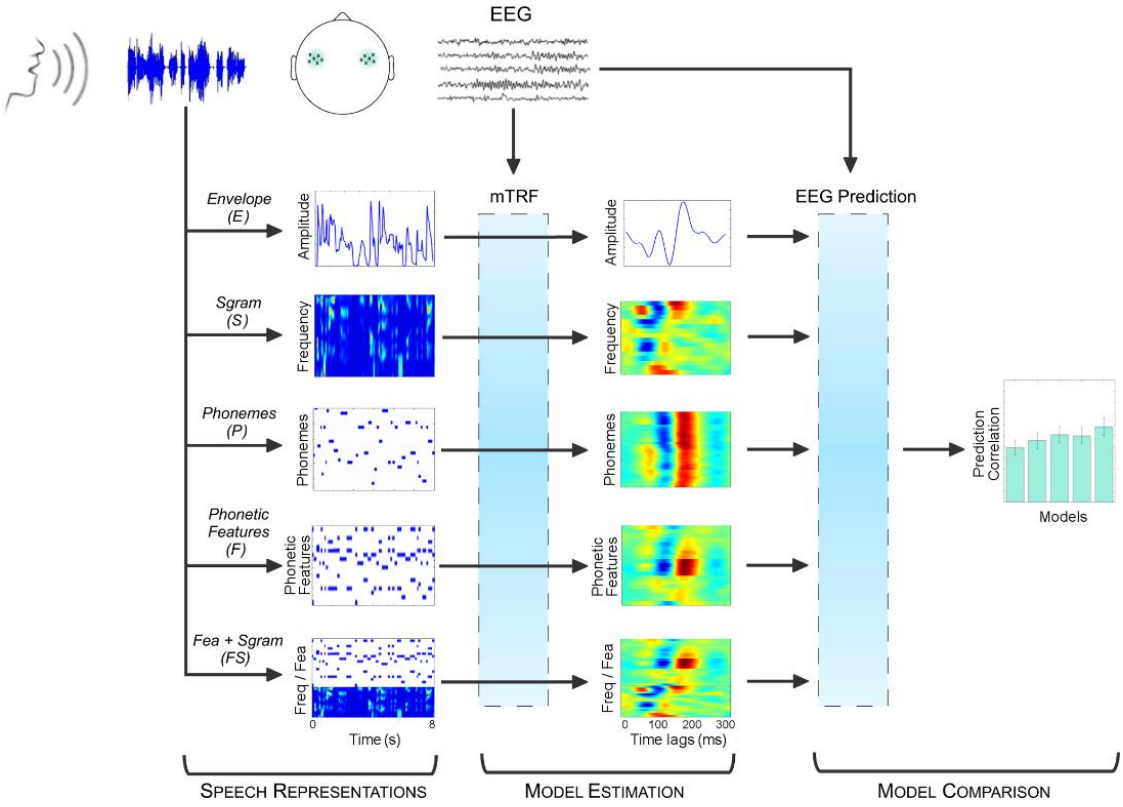


Figure 3.1: Assessing the encoding of speech features in EEG.
128-channel EEG data were recorded while subjects listened to continuous, natural speech consisting of a male speaker reading from a novel or its time-reversed complement. Linear regression was used to fit multivariate temporal response functions between the low frequency (1-15 Hz) EEG data and five different representations of the speech stimulus. Each mTRF model was then tested for its ability to predict EEG using leave-one-out cross-validation.

3.2.6 Model Evaluation

We wished to compare how each speech representation mapped to the EEG. To do this, we used a leave-one-out cross-validation approach, whereby, for each representation, an mTRF was trained on 27 trials, and used to predict the EEG data from the remaining trial. This process was repeated until the data from all trials were predicted. Prediction accuracies were evaluated by determining a correlation coefficient (Pearson's *r*) between the actual and predicted EEG data on each electrode channel. Note that silent time intervals for which the entire time-lag window had zero values were removed from the correlation evaluation because the prediction values for such silent intervals would be constant and, therefore, not meaningful when predicting EEG signals (the same intervals were removed from all speech representations). When analysing the characteristics of the mTRFs (**Figures 3.4** and **3.5**), the results were obtained using the whole dataset, with no division between training and testing data.

3.2.7 Electrode Selection

The model evaluation procedure revealed a specific distribution of EEG prediction correlations across the scalp. Importantly, there was no statistical difference in the distribution of these predictions between the 5 models (E, S, P, F, and FS; $p > 0.05$ T-ANOVA, Lehmann and Skrandies, 1980; **Figure 3.2E**). A subject-independent set of 12 electrodes from the 2 areas of the scalp with high prediction correlations was selected (6 on the left side of the scalp, and their symmetrical counterparts on the right), without biasing any of the mTRF models. Each of the selected electrodes was among the 12 electrodes with highest prediction correlations for over 90% of the cross-validation steps for every model. This subset of electrodes was used to obtain the prediction correlations presented in **Figures 3.2** and **3.3**. The average of the mTRFs across these 12 electrodes is presented in **Figure 3.4**.

3.2.8 Time-lag Selection

The presented mTRFs were first computed on a broad time-window from -150 to 450 ms. Based on visual inspection, this time interval was then restricted to lags from 0 to 250 ms as no visible response was present outside this range. This was confirmed by a quantitative search using the maximisation of the EEG prediction correlation as the objective function.

3.2.9 Multi-Dimensional Scaling

Figures 3.6 and **3.7** display the results of a multi-dimensional scaling (MDS) analysis applied to the phonemic and phonetic-features mTRF models (i.e., on the multidimensional array of weights produced by the linear regression). Given a set of 'objects', MDS is an analytic vehicle which transforms each object into a point in a multi-dimensional space. Importantly, the distances between the objects reproduce an empirical matrix of dissimilarities *Di*. In our case, the objects are phonemes (or phonetic-features) and the dissimilarities are standardised Euclidean distances between their neural responses. In particular, a phoneme object is composed of the linear regression weights at all of the 12 EEG channels of interest for each subject (channel data were not averaged in this analysis). As such, it incorporates spatial information across channels as well as temporal information across the mTRF allowing insights into the spatiotemporal activation differences between different phoneme objects. That said, the differences we report are most likely driven by temporal information (**Figure 3.2E**).

Previous research has suggested that MDS is a useful analysis tool for the categorical perception of phonetic stimuli (Shepard, 1980; Iverson and Kuhl, 1995; Chang et al., 2010), as well as other areas of research. Similar to previous research (Chang et al., 2010), we employed a non-metric MDS that minimises the mapping error measured by Kruskal Stress (Kruskal and Wish, 1978). Finally, the MDS was calculated in 5 dimensions (eigenvariates), which was enough to allow the reconstruction of the original dissimilarities with an accuracy in excess of 90% in all cases (stress ≤ 0.1) (Kruskal and Wish, 1978).

3.2.10 *F*-Scores

In a classification task, given a set of 'objects' grouped in a meaningful set of classes, the main goal of a classifier is to predict the class to which each object belongs. The *F*-Score (or *F1*-Score) is a measure of the quality of such predictions, obtained as the harmonic mean of precision and recall (Rijsbergen, 1979; Chinchor, 1992; Sasaki, 2007). The *F*-Score can be used, for example, to compare distinct classification algorithms when using the same data, to compare different conditions or, as in this case, different datasets (i.e., TRFs for particular phonemes or features) when using the same classifier.

We performed 100 repetitions of the randomised classification algorithm *k*-means (unsupervised classification, with prior knowledge of the number of classes *k*, MacQueen,

1967) to classify phonemes (and phonetic-features) and to study which classes of features are represented in the mTRFs. For each repetition, given the phonemic (or phonetic-features) mTRF model at all electrodes of interest for every subject, *k*-means returns a set of predictions for which an *F*-Score is evaluated. The values reported in **Figures 3.6** and **3.7** are the averages of these repetitions. In order to work with *k* classes with 'roughly' the same number of elements, the classification for consonants/non-consonants and for manner of articulation was performed on a revised set of classes, where the smallest classes were merged with the most similar ones.

3.2.11 Statistical Analyses

All statistical analyses were performed using a repeated measures ANOVA to compare distributions of Pearson correlation values across models and speech direction (forward and time-reversed speech) and to compare *F*-Score classifications across response intervals and speech direction. In the latter analysis, we used the jackknife method on the *F*-Scores of the 10 subjects. The values reported use the convention *F*(*df*, *dferror*). Greenhouse-Geisser corrections were made if Mauchly's test of sphericity was not met. All post-hoc model comparisons were performed using Bonferroni corrected paired *t*-tests.

While it is customary to apply Fisher's *z* transformation to Pearson correlation scores before performing statistical analysis on those scores, we did not do that for the results presented below. The rationale for the Fisher transform is to normalise the sampling distribution of the (usually skewed) Pearson's *r* values and to produce a less biased statistic. However, in our case, the *r* values are really quite low and are, generally speaking, already normally distributed. Also, it has been suggested that with large numbers of data points and small *r* values, applying a Fisher's *z* transformation can in fact lead to a more biased result (Corey et al., 1998).

3.3 Results

128-channel EEG was recorded from 10 subjects as they listened to segments of an audio-book and 10 subjects who listened to the same audio-book played in reverse (five subjects undertook both experiments). To identify neural indices of lower- and higher-level speech processing we investigated mappings between different representations of the speech and

the low frequency (1-15 Hz) EEG (**Figure 3.1A**). Specifically, we did this by using regularised linear regression to model the relationship between each speech representation and the data from each EEG channel (**Figure 3.1A**). The resulting models are commonly referred to as (multivariate) temporal response functions (Crosse et al., 2016b). We employed a cross-validation approach and mTRF models to quantify how well each speech representation related to the neural data. The quality of the prediction was assessed using correlation (Pearson's *r*).

3.3.1 Neural evidence for phonetic processing

The overarching rationale was to use variations in EEG prediction scores across speech representations as a dependent measure for assessing how well the EEG reflects the processing of lower- and higher-level speech features. Neural entrainment to speech envelopes is well established and, as such, performance of the E-model acted as a baseline with which to compare the performance of the other models. Robust mappings between speech spectrograms and high-gamma frequency ECoG have previously been shown (Pasley et al., 2012). However it is unknown whether this richer representation can be accurately indexed using low frequency EEG, something we address with the S-model. Similarly, the relationship between high frequency ECoG and a categorical phoneme representation of speech has been examined before (Mesgarani et al., 2014). However, no such relationship has been investigated for EEG (or MEG), hence the P-model. Transforming phonemes into a lower dimensional phonetic feature representation (Chomsky and Halle, 1968) frames our results in terms of the articulatory and acoustic properties of each phoneme and has advantages for the efficiency of this type of modelling. This motivated our F-model.

An important issue when considering the spectrogram representation and the phonemic/phonetic feature representations is that they are mutually highly redundant. This is because, on average, each phoneme will have a particular characteristic spectro-temporal profile. So if each phoneme were always spoken in the same way, then the two representations would be equivalent. However, in natural speech this is not the case, with significant variation in the spectro-temporal profile of a given phoneme across instances. One might thus expect that our P-model (and F-model), which is ignorant of these variations, would underperform relative to the S-model. However, it is also true that human listeners categorically perceive phonemes despite spectro-temporal variations, a

fact that is presumably underpinned by consistent neural responses to those phonemes (Okada et al., 2010; Peelle et al., 2010). Such consistent responses would be captured by our P-model, potentially leading to it outperforming the S-model, which is ignorant of the categorical nature of these utterances. Indeed, given their mutual redundancy and complementary strengths, both models may perform similarly. To attempt to reveal their complementary strengths, we also derived a model based on combining the time-aligned phonetic features and the corresponding speech spectrogram (the FS-model). Improved performance of this model over the others would suggest that the EEG is indexing the processing of both low-level acoustic fluctuations and higher-level phonetic features.

In line with this hypothesis, the average performance of the FS-model across our 12 chosen electrodes was better than all other models (ANOVA: $F(1.5,13.6) = 29.1$, $p = 2.9\text{x}10\text{-}5$; post-hoc paired *t*-test comparisons of FS with all other models: $p = 0.001$, $p = 0.002$, $p = 8.2\text{x}10\text{-}5$, $p = 0.001$ for E, S, P, and F respectively; **Figure 3.2A**). Indeed FS was best for all 10 subjects (**Figure 3.2B**). The fact that the P- and F-models are simple transformations of one another was reflected in the lack of any performance difference between them ($p > 0.05$). There was also no difference between the P- (or F-) and S-models ($p = 0.24$ and $p = 0.34$, respectively), which, as mentioned previously, was always a possibility. Importantly, we found that the model based on the envelope of speech, which has been heavily relied upon in many studies of speech neurophysiology (Luo and Poeppel, 2007; Zion Golumbic et al., 2013b; Ding et al., 2014), underperformed relative to all other models ($p < 0.01$). Furthermore, we found no lateralisation effects in the performance of any model ($p > 0.05$, ANOVA). Indeed, model performances were qualitatively similar at other scalp locations.

While we contend that the improved performance of the FS-model is evidence for the encoding of both low-level acoustic variations and higher-level phonetic features, it remained possible that this result was driven by the FS-model having more free parameters than the other models. We sought to test whether or not this was the case by investigating the performance of several other high dimensional models. Combining P and S did not outperform the FS-model ($p > 0.05$), even though it has 16 additional dimensions. Also combining the P- and F- models did not outperform either the P- or F-models alone ($p > 0.05$). These results suggest that the greater number of parameters in the FS-model does not explain our finding.

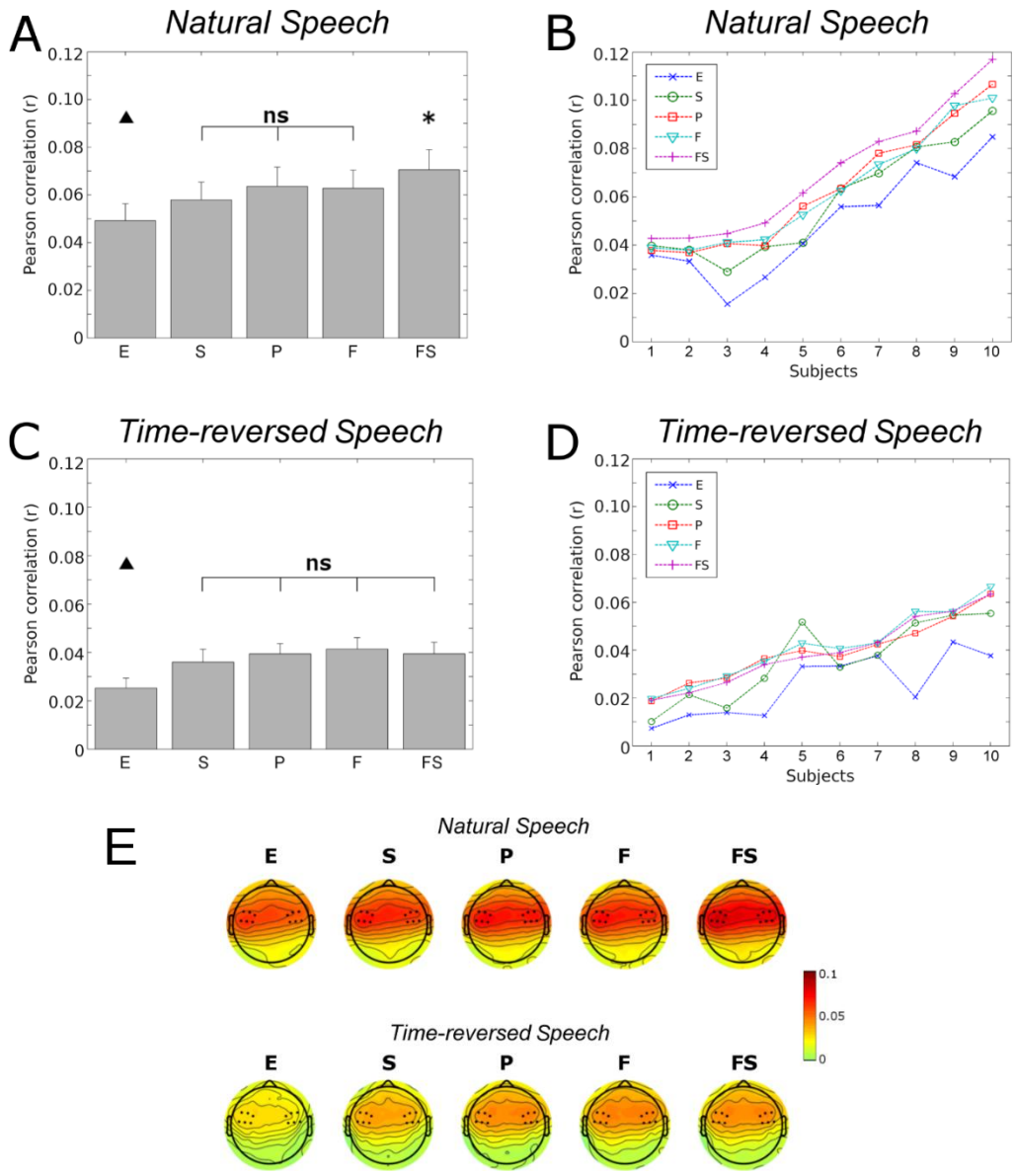


Figure 3.2: EEG responses to forward speech, but not time-reversed speech, are best predicted when speech is represented as a combination of spectro-temporal features and phonetic feature labels.
(**A**) Grand-average EEG prediction correlations (Pearson's *r*) for each speech model (mean ± SEM). While there is no statistical difference in prediction performance between the Spectrogram (S), Phonetic features (F), and Phonemic (P) models ($p > 0.05$), all of these models are better predictors of the EEG than that based on the envelope (E; ▲ $p < 0.01$). Importantly, the model based on the combination of phonetic-features and spectrogram (FS) outperforms all other models (* $p < 0.01$). (**B**) Correlation values between recorded EEG and that predicted by each mTRF model for individual subjects. The subjects are sorted according to the prediction correlations of the FS-model, which outperforms all the other models for every subject. The E-model performs worse than every other model for every subject. (**C**) Grand-average EEG prediction correlations for the time-reversed speech condition (mean ± SEM). Prediction correlations using the speech envelope are lower than those for all other models (▲ $p < 0.05$). As with normal speech, there is no statistical difference in prediction performance between the S, F, and P models ($p > 0.05$). Importantly in this case, there is also no difference between the performance of those models and that based on the combination of phonetic-features and spectrogram (FS; $p > 0.05$). (**D**) Correlation values between recorded EEG and that predicted by each mTRF model for individual subjects for time-reversed speech. The subjects are sorted according to the prediction correlations of the FS-model. The E-model performs worse than every other model for every subject. (**E**) The topographical distributions of the prediction accuracies are shown for the 5 models and for the 2 conditions: natural and time-reversed speech. A non-parametric test was performed to test for differences between these topographies (T-ANOVA). No significant differences were found ($p > 0.05$). The electrodes that were chosen for all analyses are represented by black dots.

However, to further establish the validity of our interpretation we performed the same analyses on the data from the subjects who listened to time reversed speech. Because the same speech segments were used, the same E, S, P, F and FS speech representations could be used in a time-reversed fashion. The key manipulation here is that time-reversed speech has the same long-term amplitude spectrum as natural speech, but is not perceived as intelligible speech. Overall, the prediction values were lower than for forward speech, likely as a result of differences in top-down attention and also consistent with previous research showing weaker neural entrainment to unintelligible speech (Peelle et al., 2013). However, crucially, while the E-model again performed more poorly than the others ($p < 0.05$), the FS-model in this case showed no improvement over the S-, P- or F-models ($p > 0.05$; **Figure 3.2C,D**). This supports our contention that the FS-model, in the case of forward speech, indexes the neural processing of speech features at the level of phonemes.

3.3.2 Phonetic processing across different EEG frequency bands

Given previous research positing different functional roles in speech processing for different cortical oscillations (Giraud and Poeppel, 2012) and, in particular, differential encoding of speech features by delta and theta-band entrainment (Ding and Simon, 2014), we examined the different model performances for distinct frequency bands. Phoneme level processing (i.e., FS outperforming all other models) was evident only in the delta- and theta- bands (**Figure 3.3**). The P- and F-models outperformed the S-model for the delta-band while the S-model outperformed the P- and F-models for the alpha, beta and low-gamma bands, possibly evincing the differential sensitivity of these bands to detailed acoustic information (S) and categorical phonemic processing (P and F). Relative model performances in the theta-band are qualitatively similar to the results obtained above with the broadband (1-15 Hz) signal. EEG prediction scores are very low for beta and low-gamma in keeping with the generally low SNR for EEG at these frequencies.

Given the higher prediction scores for delta, theta and alpha, and the phonetic processing effects visible using the broadband (1-15 Hz) EEG signal, we continue to analyse this broader representation of the EEG in what follows.

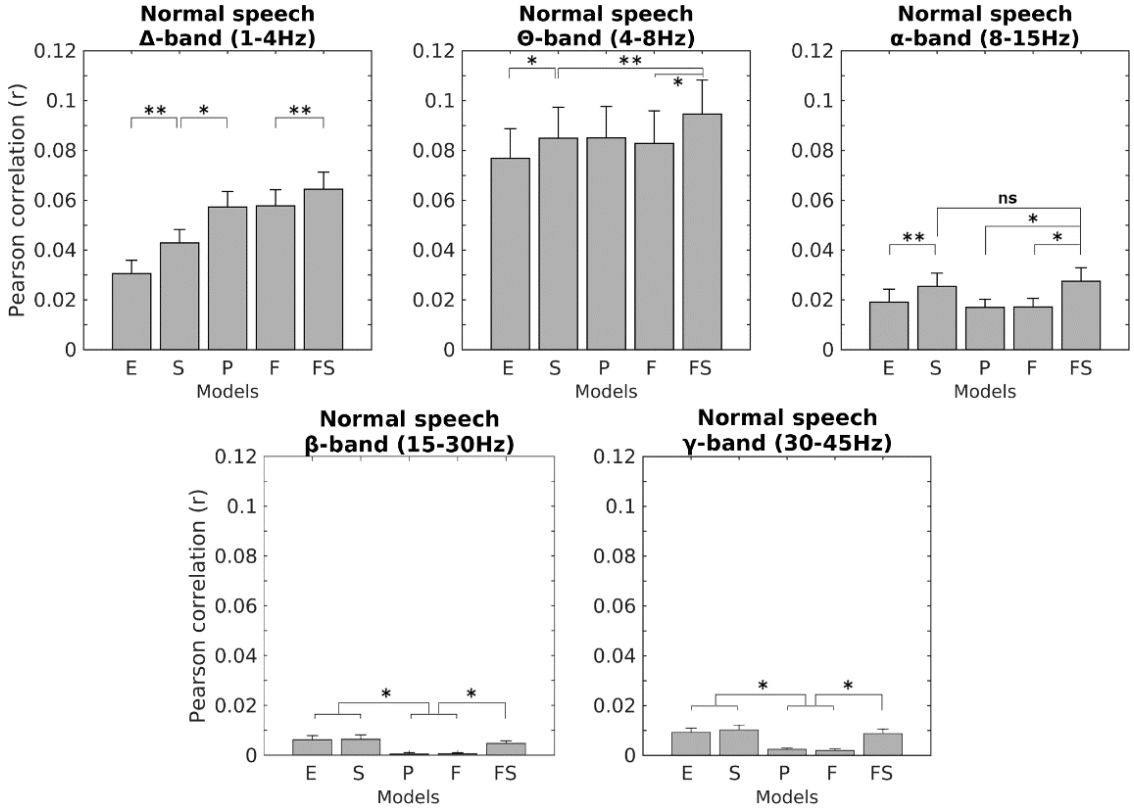


Figure 3.3: EEG response prediction for different EEG frequency bands.
Grand-average EEG prediction correlations (Pearson's *r*) for each speech model (mean ± SEM) for delta, theta, alpha, beta and low-gamma band EEG frequencies. * denotes prediction differences at the level of $p<0.05$, while ** denotes prediction differences at the level of $p<0.01$, both using planned paired *t*-tests.

The P-mTRF shows variable dynamics across phonemes (**Figure 3.4C**). To reveal groups of phonemes with similar responses, we performed hierarchical clustering on the P-mTRFs at the 12 electrodes of interest. In doing so, we found that the model could accurately discriminate consonants and vowels (32 of 35 phonemes classified correctly). For visualisation purposes (**Figure 3.4C**), we present the phonemes grouped as vowels, diphthongs, semi-vowels, and consonants and we sort within each group according to the hierarchical clustering distances. It can be seen that the mTRFs for consonants onset at around 50 ms while those for vowels do not generally show a significant response before ~100 ms. In contrast to these early differences, all phonemes show a similar response between ~150 and 200 ms. This timing pattern is consistent with previous research that has shown that acoustic-phonetic features of speech modulate activity in non-primary auditory areas from 50-100 ms onwards, with language-specific phonetic-phonological analysis starting by 100-200 ms (Salmelin, 2007). Interestingly, in the case of time-reversed speech, the P-mTRF amplitude is noticeably lower than for forward speech particularly during the 150-200 ms interval (**Figure 3.5C**).

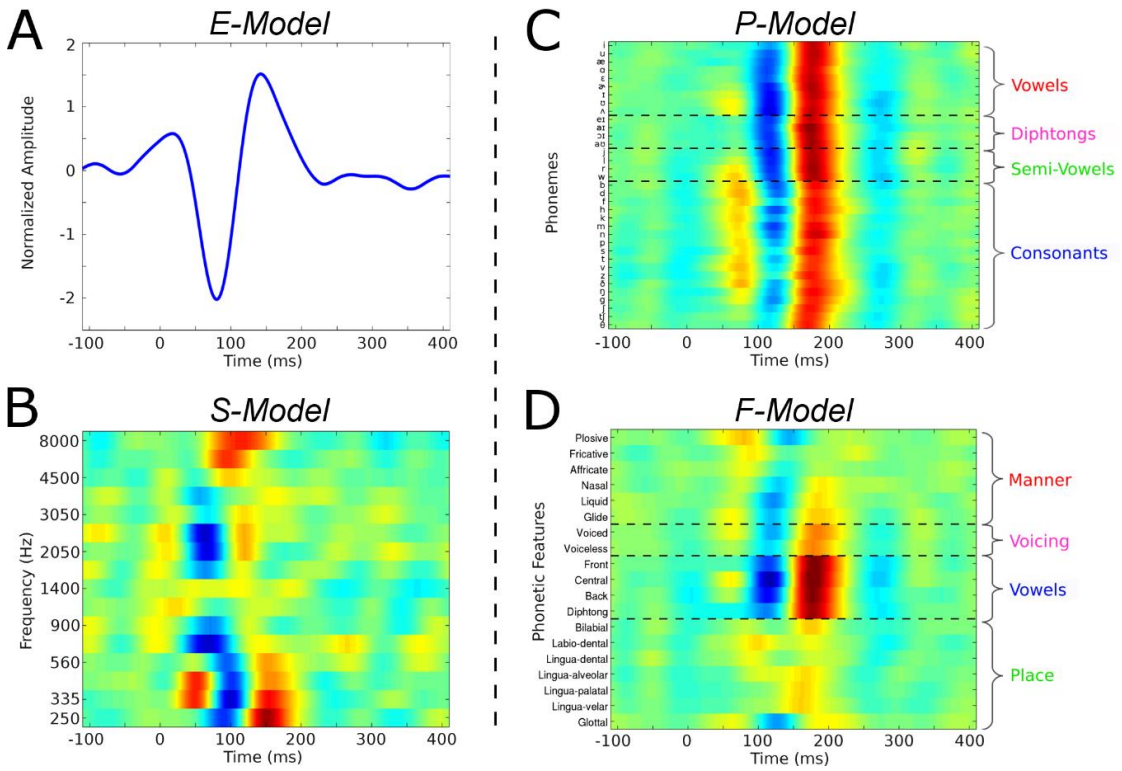


Figure 3.4: mTRF models for natural speech reflect sensitivity to different speech features. mTRFs plotted for **(A)** envelope (E), **(B)** spectrogram (S), **(C)** phoneme (P) and **(D)** phonetic feature (F) models at peri-stimulus time-lags from -100 to 400 ms for natural speech, averaged over 12 fronto-temporal electrodes. The phonemes were sorted based on a hierarchical clustering analysis on the average mTRF after grouping them into vowels, diphthongs, semi-vowels, and consonants. Horizontal dashed lines separate distinct categories of phonemes and phonetic-features.

When considering the mTRF for the F-model, it should be remembered that each phoneme is simply a combination of phonetic-features. Indeed, a linear mapping from the F-mTRF to the phonemic space produced a model that is highly correlated with the P-model ($r = 0.93$, $p = 1.6*10^{-5}$; 2-tailed *t*-test). We therefore consider these two models to be essentially equivalent. However, while the mTRF for the P-model highlights differences between vowels and consonants, the mTRF for the F-model allows us to visualise sensitivity to different articulatory speech features (**Figure 3.4D**). Again, the vowels stand out strongly from the features associated with consonants. But within each of the consonant-related features, a considerable degree of variability is evident across the specific distinctions.

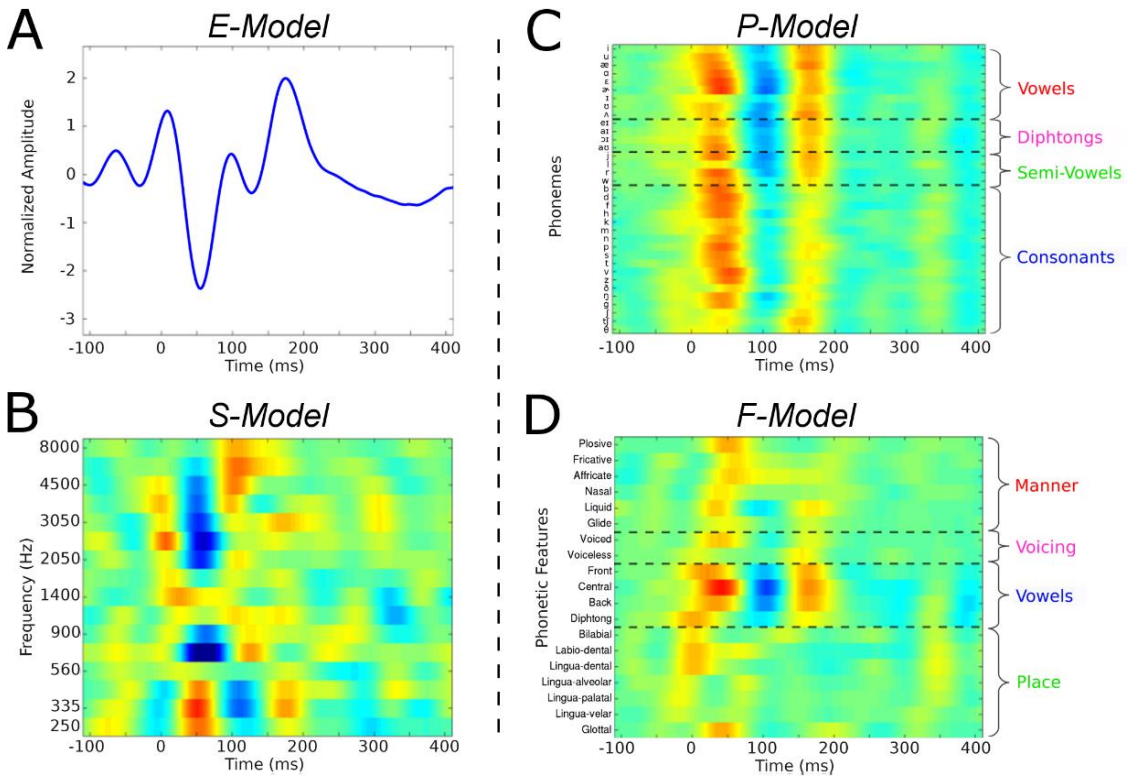


Figure 3.5: mTRF models for time-reversed speech.
mTRFs plotted for **(A)** envelope (E), **(B)** spectrogram (S), **(C)** phoneme (P) and **(D)** phonetic feature (F) models at peri-stimulus time-lags from -100 to 400 ms for natural speech, averaged over 12 fronto-temporal electrodes. The phonemes order is the same as that used in **Figure 3.4**. Horizontal dashed lines separate distinct categories of phonemes and phonetic-features.

3.3.3 Sensitivity of EEG to phonetic features as a function of latency

We wished to test the hypothesis that the sensitivity of our neural responses to different acoustic and phonetic features would increase as a function of response latency in line with what one might expect of a hierarchical system. To do this we applied unsupervised multi-dimensional scaling to the mTRFs in the following time intervals: 50 – 100 ms, 100 – 150 ms, and 150 – 200 ms, which correspond approximately to the 3 main peaks in the P-mTRF (**Figure 3.4C**). This approach allowed us to build a geometric space in which the Euclidean distance between phonemes (or phonetic-features) corresponds to the similarity of their neural responses. Furthermore, this allowed us to examine how sensitive the neural responses were to different phonetic features by quantifying how well the responses clustered according to the different groups of phonetic features that produced them. We did this by performing *k*-means clustering, where *k* is the number of groups under consideration, and then calculating the corresponding *F*-Scores (the harmonic mean of precision and recall) between the actual grouping and the result of the

clustering. All statistical tests were performed using a jackknife one-way repeated measures ANOVA with a Greenhouse-Geisser correction if the assumption of sphericity was not met.

The increasing *F*-scores as a function of latency for the P-mTRF show that the responses become more discriminative between consonants and non-consonants at longer latencies ($F(2.0, 18.0) = 3\text{x}10^9$, $p < 0.0005$; **Figure 3.6A**). Similarly, the *F*-scores for the F-mTRF show that the responses become more sensitive to different groups of phonetic features as a function of latency ($F(1.3, 11.4) = 10^5$, $p < 0.0005$; **Figure 3.6B**). Again, it can be seen that responses to vowels are clearly separable from those to consonant-related features at longer latencies. Analysis within each phonetic feature group revealed no sensitivity in our mTRFs for place of articulation, voicing or for different vowels (not shown). However the mTRFs did discriminate manner of articulation (**Figure 3.6C**), especially at longer latencies ($F(2.0, 18.0) = 215.0$, $p < 0.0005$). These results show that non-invasive neural responses to speech are sensitive to specific phonetic features and that this sensitivity increases as a function of latency.

The lack of response sensitivity to different specific vowels above, combined with the high degree of discriminability *between* vowels and consonants, caused us to wonder whether our model performance (**Figure 3.2A**) was mostly driven by this between class response sensitivity. We tested this by randomly relabelling the consonants in our time-aligned phoneme model (P) with other consonants and by relabelling the vowels with randomly chosen vowels. This led to a marked drop in EEG prediction performance (mean ± SD, $r = 0.0247 \pm 0.0009$, shuffled over 50 randomly relabelled versions of the stimulus, compared with 0.0635 for the correct P labelling). This suggests that, while the neural responses strongly discriminate between vowels and consonants, the data are also sensitive to differences within these two classes.

Finally, we repeated the above analyses for the time-reversed speech (**Figure 3.7**). In this case, consonants and non-consonants could still be discriminated in the P-mTRF ($F(1.1, 20.3) = 42.9$, $p < 0.0005$). In addition phonetic features ($F(1.3, 23.1) = 148.0$, $p < 0.0005$) and manner of articulation ($F(2.0, 36.0) = 147.7$, $p < 0.0005$) could also be discriminated. However, importantly, unlike for forward speech, there was no significant relationship between discriminability and latency for either phonetic features ($F(1.3, 11.6) = 0.1$, $p = 0.79$) or manner of articulation ($F(2.0, 18.0) = 0.46$, $p = 0.64$) and discriminability for consonants and non-consonants did not monotonically increase with latency.

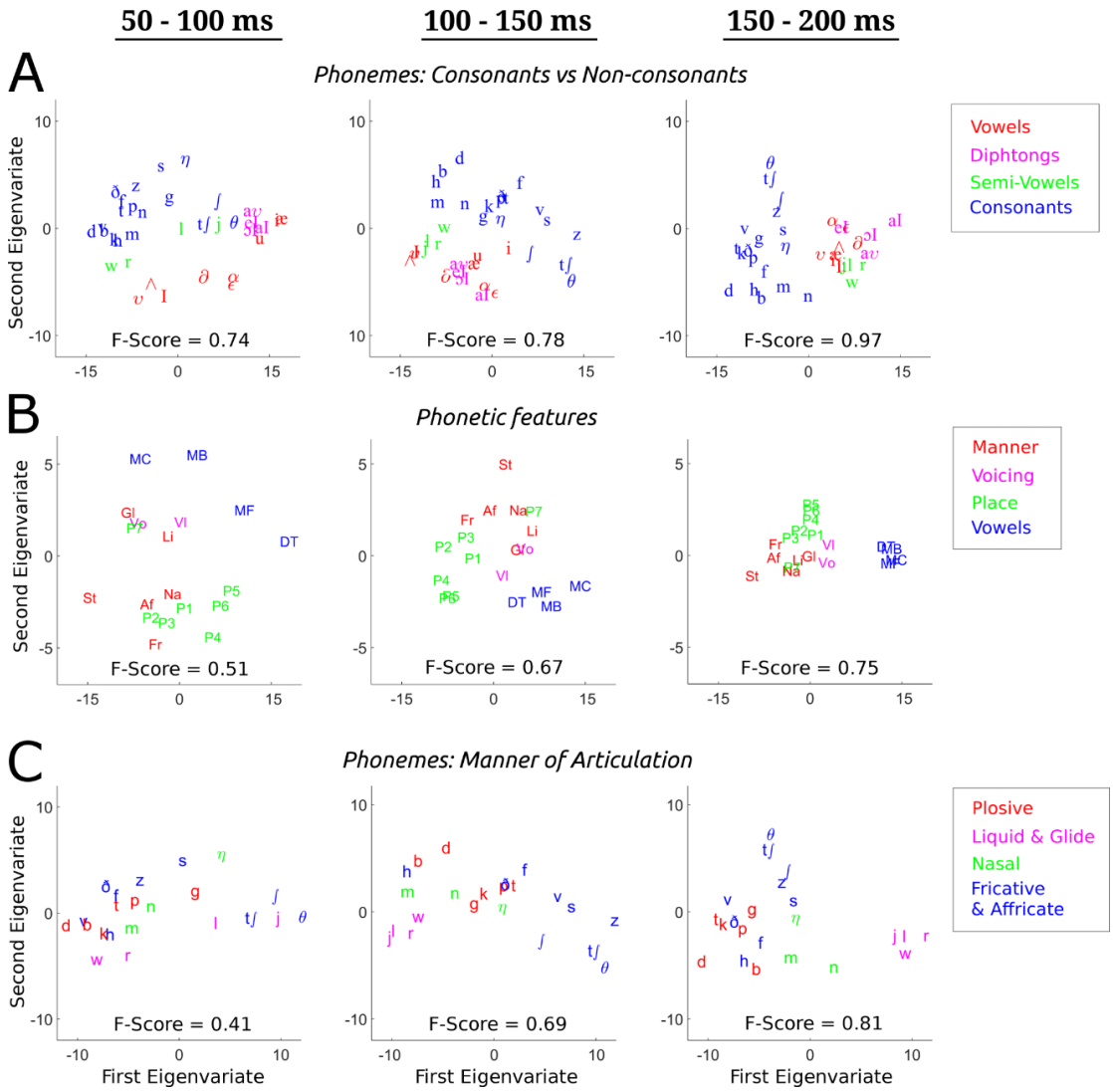


Figure 3.6: The sensitivity of EEG to speech features increases with response latency.
Multidimensional scaling (MDS) on the phonetic-features and phonemic mTRFs as a function of peri-stimulus time-lag. By carrying out repeated *k*-means classification we derive *F*-score measures that represent the discriminability of our mTRFs in each of the three time intervals 50 – 100 ms, 100 – 150 ms, and 150 – 200 ms, which correspond approximately to the 3 main peaks and troughs of the phonemic mTRF (**Figure 3.3C**). **(A)** MDS on the phonetic-features mTRF. The *F*-scores indicate the differential sensitivity of responses to manner of articulation, voicing, backness of a vowel, and place of articulation features. These *F*-scores show significant increase with response latency. **(B)** MDS on the phoneme mTRF. The *F*-scores are a measure of the binary classification of responses to consonants and non-consonants (vowels, diphthongs, and semi-vowels), which, again, significantly increases with latency. Although the classification performed was binary, the distinction between the four main classes of phonemes is evident. **(C)** MDS on the phoneme mTRF. Here the *F*-scores are computed for the four classes: plosive; liquid and glide; nasal; fricative and affricate. The four categories are progressively more separable across the three time intervals (*jackknife* method, $p < 0.0005$).

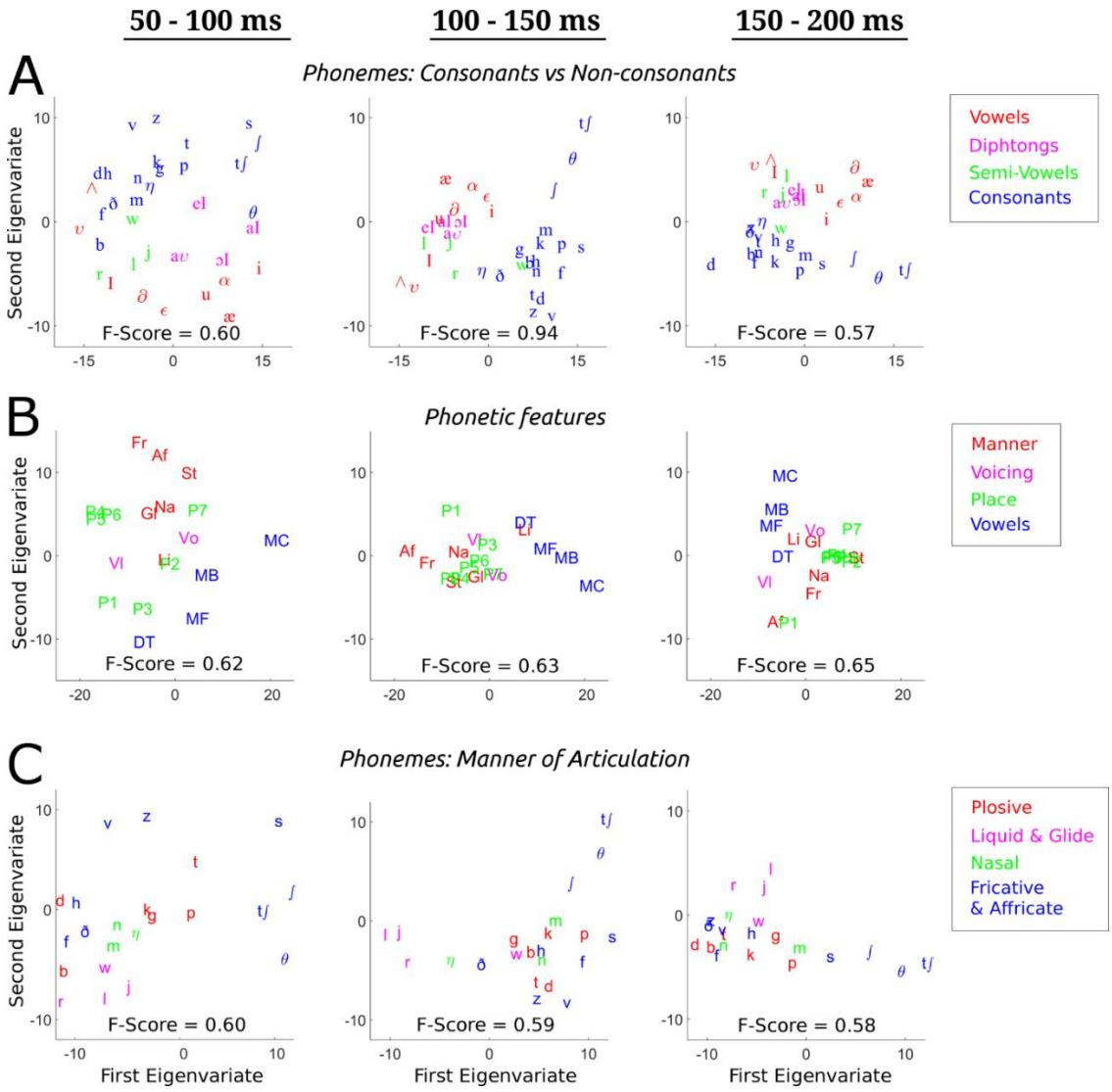


Figure 3.7: Discriminability of speech features in EEG for time-reversed speech.
Multidimensional scaling (MDS) on the phonetic-features and phoneme mTRF as a function of peri-stimulus time-lag. *F*-Scores measures are derived with the same procedure used in **Figure 3.6**. **(A)** MDS on the phonetic-features mTRF. The *F*-scores indicate the differential sensitivity of responses to manner of articulation, voicing, backness of a vowel, and place of articulation features. These *F*-scores show no significant increase with response latency. **(B)** MDS on the phoneme mTRF. The *F*-scores are a measure of the binary classification of responses to consonants and non-consonants (vowels, diphthongs, and semi-vowels). As in **Figure 3.6**, the figures show a distinction between the four main classes of phonemes. However, the same monotonic increase with latency is not seen. **(C)** MDS on the phoneme mTRF. Here, the *F*-scores are computed for the four classes: plosive; liquid and glide; nasal; fricative and affricate. There is no significant difference in *F*-score with latency.

3.4 Discussion

For humans to successfully process natural speech they must parse complex and variable acoustic inputs into categorical units and correctly encode those units as particular phonemes (Chang et al., 2010). In this chapter, in the context of natural speech, we have demonstrated that low frequency, noninvasively recorded EEG indexes this categorical

phonemic level processing. Furthermore, we have shown that the articulatory features of speech can be best discriminated by responses at longer latencies, in line with what one might expect of a hierarchical system.

3.4.1 EEG measures of cortical entrainment reflect speech-specific processing

Our findings have important implications for current theories on cortical entrainment to the envelope of speech (Luo and Poeppel, 2007; Aiken and Picton, 2008; Giraud and Poeppel, 2012; Ding and Simon, 2014). In particular, we have shown that the processing of different speech features that covary with the envelope can be dissociated according to the neural responses they elicit. Therefore, neural measures based on the envelope alone are likely to include contributions from neural populations at different levels of the speech processing hierarchy. Given the relatively modest difference in modelling performance between our FS- and S-models, it is entirely possible that the speech-specific contribution to measures of cortical entrainment is relatively small in comparison to the more general response to the stimulus acoustics. As we know from Chapter 2, one brain region that could be responsible for such a contribution is the superior temporal sulcus. It has been suggested that STS is involved in phonological-level processing bilaterally (Hickok and Poeppel, 2007), a finding that fits with the lack of any lateralisation effects in our prediction performances. While this region has been implicated in many other cognitive domains (Hein and Knight, 2008), recent neuroimaging work has suggested that it may represent a special locus of speech analysis that is distinct from lexical, semantic or syntactic processes (Overath et al., 2015). The notion that speech-specific effects in EEG may derive from a relatively small contribution from a specific brain region such as STS would partly explain why it has been so difficult to definitively say whether or not envelope entrainment measures reflect anything more than low-level processing of the acoustics of speech (Ding and Simon, 2014).

Recently it has been suggested that there may be different functional roles for entrainment at different frequencies with theta-band entrainment (4-8 Hz) encoding speech features critical for intelligibility and delta-band entrainment (1-4 Hz) being related to the perceived, non-speech-specific acoustic rhythm (Ding and Simon, 2014). Our finding that FS outperforms all other models for delta- and theta-bands (**Figure 3.3**) suggests that both of these bands may reflect speech-specific processing. One attempt to reconcile these views is to suggest that relying on envelope tracking as a dependent

measure, particularly for delta-band where it performs poorly, results in a lower sensitivity to subtle speech-specific effects. While we have argued that these speech-specific effects can be seen by comparing FS and S performances, it is also worth noting how the S- and F-model performances differentially vary across frequency bands (e.g., compare the relative model performances for delta and alpha). This variation potentially provides another way to disambiguate lower- and higher-level speech processing effects, something that is investigated in the next chapters.

3.4.2 A novel approach for studying natural speech perception: Further requirements and potential impact

Our finding aligns well with recent invasive ECoG research investigating the encoding of natural speech in the human brain (Chang et al., 2010; Mesgarani et al., 2014). Specifically, based on recordings from the superior temporal gyrus in epilepsy patients, high gamma frequency (75-150 Hz) activity was shown to encode an acoustic-phonetic representation of speech. Based on this, it has been suggested that the STG may be a transitional stage in the auditory processing hierarchy, early enough to still encode the acoustic features of speech but high enough to exhibit response selectivity to complex spectro-temporal patterns (Shamma, 2014). The fact that the ECoG recordings were shown to be optimally sensitive to intermediate acoustic-phonetic speech features at an intermediate response time lag of around 150 ms (Mesgarani et al., 2014), agrees reasonably well with the increased discriminative power of our EEG responses at this latency. While we have speculated that our findings may have specific contributions from STS, the concordance with ECoG from STG suggests that the analysis framework we have outlined may represent an important mechanism for applying findings from the ECoG community into research with a wider variety of subjects including infants (Kuhl, 2010), children with developmental difficulties (Gervais et al., 2004), the elderly (Ruggles et al., 2012) and patients with psychiatric disorders (Li et al., 2009). This is particularly important because much of the EEG/MEG research in these cohorts relies on stimuli composed of discrete syllables, leading to a literature that is very limited in what it can say about the parsing and processing of continuous speech (e.g. Kuhl, 2010).

Developing further insights using our approach would benefit from an ability to disentangle the activity from the many neural sources that are concurrently active during speech processing. While this issue is often seen as a shortcoming of EEG and MEG, it can also be seen as a strength in terms of the global view of hierarchical processing that

these methods provide. But it will still be necessary to further characterise how different speech representations map to different neural responses and to determine which specific neural populations are responsible for those responses. Furthermore, it will be necessary to disentangle how much the cortical entrainment of speech is driven by additive evoked activity and how much by the entrainment of ongoing oscillations (Schroeder and Lakatos, 2009; Giraud and Poeppel, 2012; Zion Golumbic et al., 2013b). One potentially fruitful approach to address these questions is to manipulate the relative amount of low- and high-level information that is available in the speech stimuli with a view towards disambiguating the information contained within our S- and F-models (e.g. Zoefel and VanRullen, 2015). Indeed this is already possible to an extent by considering the difference between the FS- and S-model performances, which we contend is likely to reflect phonemic level processing in relative isolation. Importantly this difference was positive for each and every subject. As such, it has the potential to act as a dependent measure in research aimed at understanding speech processing in particular populations. The sensitivity of response functions to different phonetic features, and how that sensitivity varies with latency also represent potentially useful dependent measures of speech-specific processing. These questions and others are at the core of the following chapters.

3.5 Summary

This chapter introduced a novel framework to investigate speech perception using non-invasive electroencephalography (EEG) and an experimental paradigm based on natural speech perception. Using this approach, we provided evidence for categorical phoneme-level speech processing by showing that the relationship between continuous speech and neural activity is best described when that speech is represented using both low-level spectro-temporal information and categorical labelling of phonetic features. Furthermore, the mapping between phonemes and EEG becomes more discriminative for phonetic features at longer latencies, in line with what one might expect from a hierarchical system. Importantly, these effects are not seen for time-reversed speech.

Chapter 4. Isolating neural indices of continuous speech processing at the phonetic level

4.1 Introduction

Chapter 3 introduced a novel approach to non-invasively index the cortical processing of natural speech at the level of phonemes. Here, we investigate the possibility to use this procedure to extract quantitative indices of phonological processing in relative isolation. The previous chapter suggested that the contrast between the measures of EEG predictability for the FS- and S-models (FS-S) may represent such an index. This study aims to test the link between this measure and phonological perception. In doing so, we also expected our measures to provide new insights for one of the major unresolved questions in speech perception: what are the cortical mechanisms underpinning the integration of prior knowledge with sensory input?

As detailed in Chapter 2, successful speech comprehension in noisy, real-world environments is carried out by a complex hierarchical system in the human brain (Chang et al., 2010; Okada et al., 2010; Peelle et al., 2010; DeWitt and Rauschecker, 2012; Hickok and Small, 2015). In such cases, it is widely acknowledged that an active cognitive process takes place where speech perception is strongly influenced by prior knowledge and a contextual expectation of upcoming speech input (McClelland and Elman, 1986; Davis and Johnsrude, 2007; McClelland, 2013; Heald and Nusbaum, 2014; Leonard and Chang, 2014). However the nature of this influence is not yet well understood.

Firstly, it remains unclear at what hierarchical processing stages – and in particular how early – the encoding of speech is affected by top-down influence (Davis and

Johnsrude, 2007). Studies using prior information to enhance the perception of degraded speech report that subjects experience a strong perceptual pop-out effect, whereby they report a marked increase in the perceived clarity of the speech as they process it in real-time (Blank and Davis, 2016; Holdgraf et al., 2016; Tuennerhoff and Noppeney, 2016). This suggests that prior information might affect speech processing *in situ* in lower-level sensory processing areas at the acoustic and phonetic encoding stages, something that has been observed for effects such as phoneme restoration in noise (Leonard et al., 2016). However, event-related potential evidence on this issue has suggested that prior information first modulates activity in higher-order areas which then feeds back to affect lower-level sensory processing at longer latencies (Sohoglu et al., 2012).

A second unresolved issue is the mechanism through which prior information affects bottom-up sensory processing. As we know from Chapter 2, one view is that the neural encoding of a stimulus is enhanced by expectation (sharpening theories) (McClelland and Elman, 1986; Mirman et al., 2006). The predictive coding theory, instead, proposes that discrepancies (or errors) between what is predicted and what is received are passed from one level to the next within the speech processing hierarchy (Friston, 2005; Arnal and Giraud, 2012; Giraud and Poeppel, 2012).

This chapter aims to confirm the effectiveness of our novel approach at isolating neural indices of speech-specific processing by providing qualitative and quantitative ways of disentangling contributions from acoustic- and phoneme-level neural processing. At the same time, the ability to disambiguate contributions reflecting the processing of low-level speech acoustics from those reflecting the processing of categorical phonetic features (Mesgarani et al., 2014; Di Liberto et al., 2015) provides a new way to investigate the underpinnings of speech comprehension. In particular, this study examines these two issues: 1) how early in the hierarchy is speech encoding affected by prior information, and 2) is the increase in perceived clarity that comes with prior information reflected in an enhancement or suppression of activity at particular hierarchical stages. Here, we use these models in an attempt to shed more light on how prior information affects speech encoding by manipulating the perceived clarity of degraded speech stimuli. Specifically, we hypothesised an increase in the strength of phonetic-feature encoding between the cases where subjects hear unintelligible degraded speech versus when they can understand that same degraded speech as a result of having prior information.

The findings described in this chapter were presented at several international conferences, were published as a conference paper: “Isolating neural indices of

continuous speech processing at the phonetic level", *Advances in Experimental Medicine and Biology*, 894: 337-345, April 2016, and have been submitted as a research article: "Cortical measures of phoneme-level speech encoding correlate with the perceived clarity of natural speech", *in review*.

4.2 Methods

4.2.1 Subjects and Data Acquisition

Fourteen healthy subjects (8 males, aged between 21 and 31) participated in this study. Electroencephalographic (EEG) data were recorded from 128 electrode positions (plus 2 mastoid channels). Data were filtered over the range 0 – 134 Hz and digitised with a sampling frequency of 512Hz using a BioSemi Active Two system. Monophonic audio stimuli were presented at a sampling rate of 44,100 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 on a crosshair centred on the screen, and to minimise motor activities for the duration of each trial. The study was undertaken in accordance with the Declaration of Helsinki and was approved by the Ethics Committee of the School of Psychology at Trinity College Dublin. Each subject provided written informed consent. Subjects reported no history of hearing impairment or neurological disorder.

4.2.2 Stimuli and Experimental Procedure

Audio-book versions of two classic works of fiction read in American English by the same male speaker were partitioned into 10-second speech snippets using MATLAB software (The MathWorks Inc.). 120 snippets were randomly selected for the experiment. In order to alter the intelligibility of the speech, a method known as noise-vocoding was implemented (Shannon et al., 1995; Davis and Johnsrude, 2003). This method filters the speech into a number of frequency-bands and uses the amplitude envelope of each band to modulate band-limited noise. Specifically, the speech for this experiment was vocoded using three frequency-bands logarithmically spaced between 70 and 5000Hz according to Greenwood's equation (70 – 494 – 1680 – 5000 Hz) (Greenwood, 1961).

Each EEG standard trial consisted of the presentation of 3 speech segments (**Figure 4.1A**). The first segment (NP: no prior knowledge) was degraded using noise-vocoding;

the second one (C: clear) was the same 10-second speech segment, but in its original clear form; and the third presentation (P: prior knowledge) was the noise-vocoded version again. As such, the first (NP) and third (P) speech segments involved identical acoustic stimuli, but it was hoped that the perceived clarity of the third segment (P) would be improved by the prior information provided by the interleaved segment C (perceptual pop-out effect, Section 2.2.1). As a control measure, we also included deviant trials. These trials consisted of a modified version of NP and/or P, where a random segment of ~5 seconds was replaced with words from a different trial. For both NP and P, the probability of a deviant stimulus was set to 10%.

Participants were asked to make two judgements based on the stimuli. First, after presentation of segment C, they were asked to decide whether the first vocoded segment, NP, was deviant (different from C) or standard (the same as C). And second, after presentation of the second vocoded segment, P, they were asked to decide whether it was a deviant (different from C) or standard (the same as C). More specifically, they were asked to make both of these decisions using a level of confidence from 1 to 5 ('definitely a deviant', 'probably a deviant', 'I don't know', 'probably a standard', and 'definitely a standard'). For standard trials, a higher confidence level when comparing segments P and C than when comparing segments NP and C was taken as evidence of enhanced perceived speech clarity. This score was normalised by subtracting a subject-specific baseline that was obtained by performing the same operation on deviant trials (see Section 4.3 for a better understanding of the rationale behind this normalisation).

Prior to the taking part in the full experiment the participants were presented with a number of noise-vocoded speech snippets for approximately 10 minutes. The goal of this was to enable subjects to become familiar with the peculiarity of noise-vocoded speech without allowing so much exposure as to enable substantial perceptual learning to take place (Sohoglu and Davis, 2016).

4.2.3 Stimulus Representations

Following Chapter 3, mTRFs were estimated based on distinct representations of the speech stimulus:

1. Spectrogram (**S**): This was obtained by partitioning the speech signal into three frequency-bands logarithmically spaced between 70 and 5000 Hz according to Greenwood's equation (70 - 494 - 1680 - 5000 Hz, the same used for the vocoder)

(Greenwood, 1961), and computing the amplitude envelope for each band, as in Equation 3.2.

2. Phonetic-features (**F**): As in Chapter 3, with the difference that only 18 phonetic features were used. Specifically, the feature 'affricate' was excluded. In fact, because of the use of short trials, it had no occurrences in several trials.
3. Finally, we propose a model that combines F and S (**FS**).

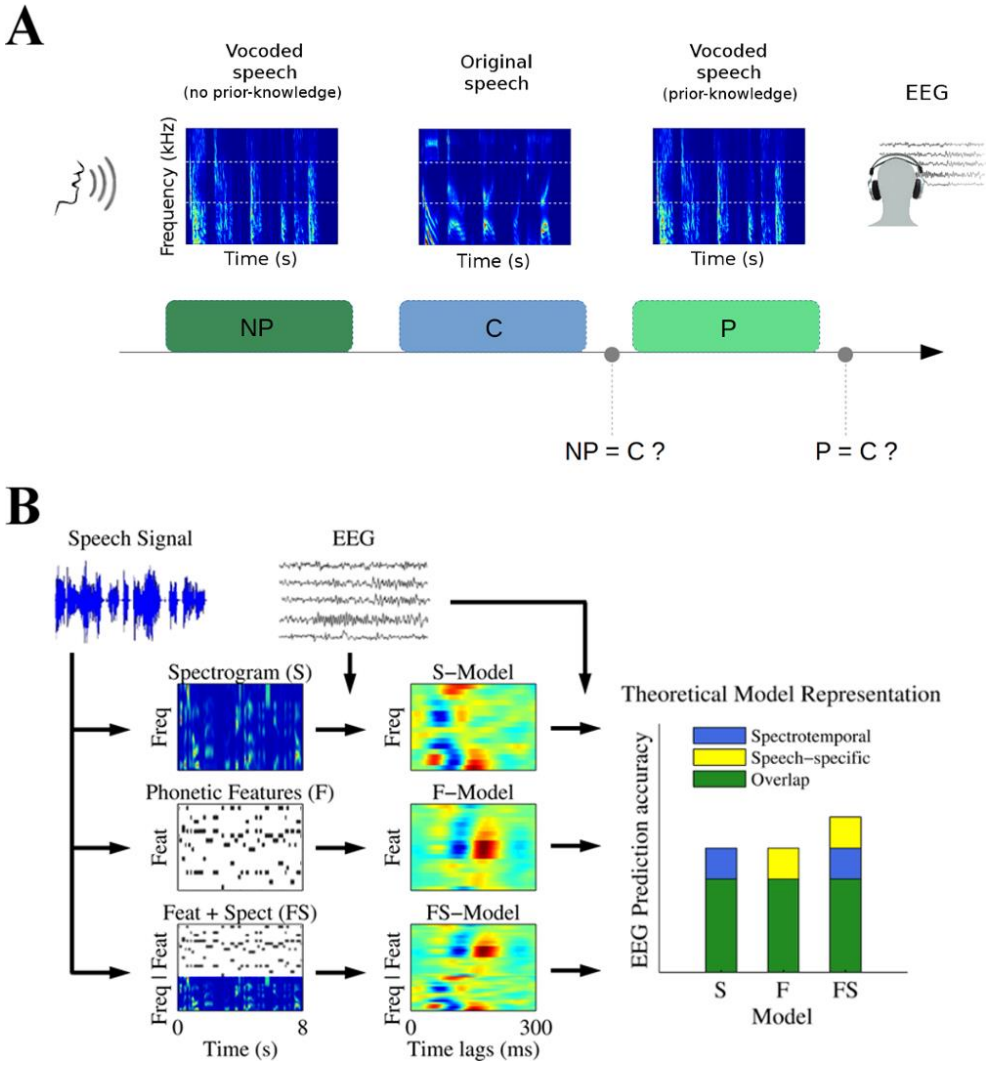


Figure 4.1: A pop-out experiment to modulate speech perception.
(A) Experimental setup. EEG data were recorded while subjects listened to groups of three 10-s long speech snippets. In standard trials, the first (NP: no prior knowledge) and the third (P: prior knowledge) speech snippets were a three-channel noise-vocoded version of the second snippet (C: clear). In deviant trials, either the first or third snippets (or both) did not fully match the second snippet. After C and P, participants were asked to identify the first and the second vocoded snippets respectively as matching the clean speech or not (i.e., standard or deviant trial). **(B)** Analysis approach. Linear regression was used to derive mappings from different speech representations to the EEG. Regression models were fit for the acoustic spectrogram (S), a set of time-aligned phonetic features (F), and a combination of the two (FS). Each model was then tested for its ability to predict the EEG using leave-one-out cross-validation. The information reflected by the S- and F-models were suggested to be largely overlapping. However, each of them carry some contributions that are unique of spectrotemporal and speech-specific (phoneme-level) processing respectively. Hence, the combined FS-model outperforms the two individual models as it combines the information that both of them encode.

Based on the above three representations, we have also previously suggested that one can attempt to isolate the *unique* contribution that derives from phonetic-feature level processing by subtracting the performance of the S-model from that of the FS-model (i.e., **FS–S**; Di Liberto et al., 2015; Di Liberto and Lalor, 2017).

A couple of final notes on our stimulus representations. Below, we also used a univariate envelope representation of the speech (E) for visualisation purposes. This was calculated as the sum of the three band-limited envelopes that constitute the S representation. Chapter 3 also included a phonemic representation of the speech (a multivariate time-series of forced aligned phonemes, similar to F). However, because of the limited amount of speech data used in the present study, rarer phonemes did not have a sufficient number of occurrences to produce a good model fit. As a result, we did not include this representation in the present study and focused our analysis on the more fundamental phonetic-features model. As an aside, if it were of interest, the scalp responses to phonemes can still be visualised by performing a linear projection of the F-model (in fact, a phoneme can be represented as a combination of specific phonetic features). Please refer to Section 3.2.2 for a more detailed description of these speech representations.

4.2.4 EEG Data Analysis

The EEG signals were analysed offline using MATLAB software. Because of suggestions that speech entrainment in the delta- (1–4 Hz) and theta-bands (4–8 Hz) might have different functional roles in speech processing (Ding and Simon, 2014), we analysed these two EEG bands separately. Specifically, the data were digitally filtered into the two frequency-bands of interest using Chebyshev Type-2 band-pass filters with pass-band in delta-band and in theta-band. Next, signals were down-sampled to 128 Hz, and referenced to the average of the two mastoid channels. EEG channels whose time-series data had a variance that exceeded three times that of the surrounding channels were identified as being excessively noisy. And the data on those channels were replaced by spline interpolating the data from the surrounding clean channels using EEGLAB software (Delorme and Makeig, 2004).

As performed in Chapter 3, a forward mTRF analysis was conducted to quantify how well the EEG reflects the encoding of the various speech representations. Regularised linear regression was used to create a mapping between the EEG and the

three abovementioned speech stimulus representations (**Figure 4.1B**). Speech stimuli and the corresponding EEG responses were partitioned into 10 equal-sized subsets S_1, S_2, …, S_{10}, and R_1, R_2, …, R_{10} respectively. 10-fold cross-validation was employed on these partitions to compare how each speech representation (S, F, and FS) mapped to the EEG. In particular, EEG signals of a subset i (R_i) were predicted using each distinct speech representation model fit on all the left-out partitions (1,…,i-1,i+1,…,10), and prediction accuracies were quantified for each electrode using a Pearson correlation. To optimise performances within each speech representation model, we conducted a parameter search (over the range 10^{-3}, 10^{-2}, …, 10^{5}) for the regularisation parameter λ within each speech representation model. This procedure maximised the EEG prediction accuracy averaged across trials, subjects, and all 128 electrodes. The combination of regularisation and cross-validation controlled for overfitting and prevented bias toward the test data used for quantifying the prediction accuracies (Crosse et al., 2016b).

The time-lag window was set between -50 and 250 ms, as the most accurate EEG prediction accuracies for clear speech were in this window. After the optimisation of the regularisation term λ and of the time-lag window, the set of 12 consistently well-predicted electrodes identified in Chapter 3 was selected for calculating the EEG prediction accuracies.

4.2.5 Statistical Analysis

Statistical analyses were performed using a repeated measures ANOVA to compare distributions of Pearson's correlation values across models. ANOVA analyses were conducted after verifying that the normality assumption was not violated, which was assessed both visually (QQ plots; not shown) and quantitatively (Shapiro-Wilk test). The values reported use the convention $F(df, dferror)$. Greenhouse-Geisser corrections were made if the assumption of sphericity was not met (as indicated by a significant Mauchly's test). All post hoc model comparisons were performed using Bonferroni corrected paired t-tests. Two-tailed permutation tests with 200,000 repetitions were used for pair-wise comparisons if the assumption of normality was violated (Shapiro-Wilk test).

4.3 Results

4.3.1 Prior knowledge enhances perceived speech clarity

Participants were asked to identify the first (NP) and the second (P) speech vocoded streams as a standard (*St*) or deviant (*D*) presentation using a level of confidence from 1 to 5 (from 'definitely a deviant' to 'definitely a standard' respectively). The response distribution for each condition (averaged across subjects; **Figure 4.2A**) indicates that participants were more confident in identifying standard trials when prior knowledge was available (St_P compared to St_{NP}), while this was not the case for deviant trials (D_P compared to D_{NP}). Note that subjects were instructed to report detection of a deviant trial only if they heard a difference with the corresponding clear speech snippet. But because perceptual pop-out did not occur for the modified portion of the D_P trials, this was a more difficult determination for subjects to make. For this reason, prior knowledge improved the standard but not the deviant detection scores.

A significant enhancement of the detection score from NP to P was observed for standard trials ($St_P > St_{NP}$, permutation test, $p = 0.001$), which confirms that prior knowledge had an effect on the subjects' confidence in detecting standard trials. However, this alone is not sufficient to draw conclusions about the effects of prior knowledge on the perceived speech clarity. This is because it was possible that subjects may have been biased to respond to both standard and deviant stimuli as standard trials when prior information was available. For example, this was the case for subject 12, whose individual behavioural scores are reported in **Figure 4.2B** (bottom panel). In contrast, subject 5 (**Figure 4.2B**, top) exhibited an increase of speech clarity with prior knowledge, as detection for both standard and deviant improved for P trials. In order to control for such biases across individual subjects, a subject-specific baseline was derived using deviant trials and subtracted from the confidence level for standard trials. This corrected behavioural measure ($St - D$) exhibited a significant interaction with prior knowledge ($St_P - St_{NP} > D_P - D_{NP}$, permutation test, $p = 10^{-6}$). This result, which is depicted in **Figure 4.2C**, indicates an increase in perceived speech clarity due to prior knowledge of the upcoming stimulus. This perceptual enhancement can be summarised for each single subject using the following quantitative measure:

$$\Delta_{\text{Clarity}} = (\text{St}_\text{P} - \text{St}_\text{NP}) - (\text{D}_\text{P} - \text{D}_\text{NP}). \qquad (4.1)$$

Interestingly, the result in **Figure 4.2** shows that the NP vocoded speech snippets, although severely degraded, were perceived as partially intelligible rather than completely unintelligible ($St_{NP} > D_{NP}$, permutation test, $p = 10^{-6}$). These results indicate that, as hypothesised, prior information led to clearer perception of the noise-vocoded speech stimuli, a perceptual difference that we have quantified as Δ_{Clarity}.

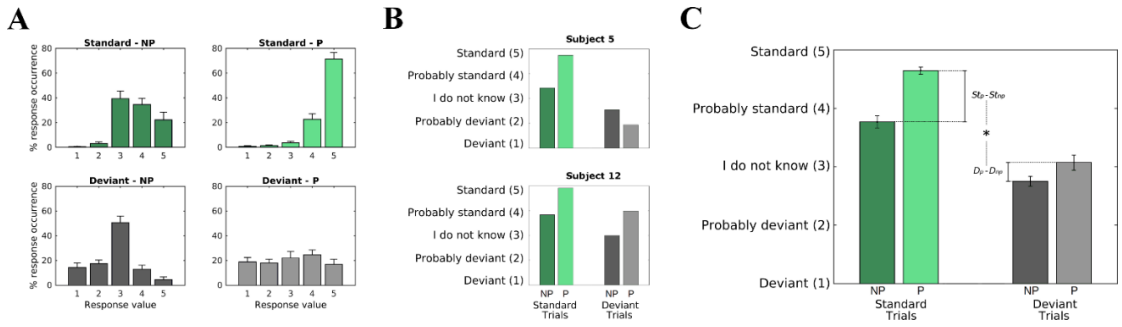


Figure 4.2: A behavioural measure of speech clarity reflects the effect of prior knowledge.
Subjects were presented with sequences of vocoded-original-vocoded speech snippets and were asked to identify the two noise-vocoded streams (NP and P stimuli) as standard or deviant presentations by comparing them with the original speech snippet. Responses consisted of a level of confidence from 1 ('Definitely a deviant') to 5 ('Definitely a standard'). **(A)** The response distributions (mean percent occurrence ± SEM) confirm that subjects were more confident in detecting standard trials when prior knowledge was available. **(B)** The confidence level for two selected subjects. The result in the top panel shows that subject 5 improved in detecting both standard and deviant trials when prior knowledge was available, which we interpret as evidence for an increase in perceptual clarity. In contrast, subject 12 (bottom panel), responded with higher values to P stimuli for both standard and deviant trials. In this case, the positive St_P-St_{NP} cannot be assumed to purely reflect an increase in perceived clarity, as deviants were not detected. **(C)** The confidence level averaged across all subjects (mean ± SEM) is here reported for NP and P stimuli, and for both standard and deviant trials. The increase in confidence due to prior knowledge is larger for standard than for deviant trials (*$p < 0.05$).

4.3.2 Dual effect of prior knowledge on the cortical entrainment to speech features

EEG predictability measures were derived using a forward mTRF model that estimates an optimal linear mapping from a speech representation to the corresponding scalp-recorded EEG signal. These predictability measures were derived for different frequency bands (delta, theta) and models (S, F, and FS). A significant interaction between these two factors emerged from a unified 2 x 3 ANOVA analysis for the C and NP conditions, but not for P (two-way ANOVA, C: $F(1.37,17.85) = 6.261$, $p = 0.015$, effect size = 0.33; NP: $F(1.19,15.48) = 8.454$, $p = 0.008$, effect size = 0.39; P: $F(1.26,16.42) = 0.233$, $p = 0.692$, effect size = 0.018). Based on this interaction, follow up one-way ANOVAs were conducted for the delta- and theta-bands separately and the results were compared between the no prior knowledge (NP), clear speech (C), and prior knowledge (P) stimuli. In the delta-band, the analysis for C stimuli (**Figure 4.3A, top**) showed that the combined

FS-model performed better than both S- and F-models, and that the F-model performed better than the S-model (ANOVA: $F(1.41,19.70) = 48.226$, $p = 1.7*10^{-7}$; post hoc paired *t*-test comparisons: $p = 10^{-6}$, $p = 3.5*10^{-5}$, $p = 9*10^{-4}$ for S vs FS, F vs FS, and S vs F respectively). Furthermore, the analysis for C stimuli in theta-band (**Figure 4.3A, bottom**) showed that the combined FS-model performed better than both S- and F-models, however no significant difference emerged between the F-model and the S-model (ANOVA: $F(1.26,16.37) = 14.490$, $p = 8.5*10^{-4}$; post hoc paired *t*-test comparisons: $p = 0.002$, $p = 5*10^{-6}$, $p = 1$ for S vs FS, F vs FS, and S vs F respectively). These results are consistent with those obtained for clear natural speech using a different dataset in the previous chapter.

Chapter 3 also suggested that isolated indices of speech-specific processing could be quantified using this analysis framework. In particular, as depicted in **Figure 4.1B**, it was suggested that this could be done by noting that the FS-model is sensitive to activity reflecting the processing of both sound acoustics and categorical phonetic features, while the S-model does not explicitly encode phonetic features and should thus be less sensitive to the categorical processing of those features. Therefore, it was proposed that any difference in EEG prediction accuracy between the two models would be due to the fact that the FS-model captures extra activity reflecting the processing of categorical phonetic features. And, as such, it was suggested that one could isolate a measure of speech-specific cortical processing at this level by subtracting r_S from r_{FS}. Here, we hypothesised that this measure would be specifically sensitive to differences in perceived clarity as a result of prior knowledge. Specifically, our hypothesis was that, because the perceived speech clarity (and therefore intelligibility) of the two conditions differed as a result of prior knowledge, we would see a clear increase in our proposed isolated measure of phonetic feature-level processing (FS-S) with prior knowledge. In line with other work (Holdgraf et al., 2016), we also wished to explore possibility that top-down effects on the processing of speech may impact even earlier stages of speech encoding at the level of acoustics, as indexed via the S-model. The effect of prior knowledge on this FS-S measure was quantified as:

$$\Delta(\text{FS-S}) = (r_{FS} - r_S)_P - (r_{FS} - r_S)_{NP}. \quad (4.2)$$

In line with our hypothesis, we found that Δ(FS-S) in the delta-band was positively correlated with the behavioural measure $\Delta_{Clarity}$ across subjects (**Figure 4.3B**). That is to say, the larger the enhancement in speech clarity due to prior information for a given

subject, the bigger Δ(FS-S) for that subject (Pearson's correlation coefficient $r = 0.63$, $p = 0.015$). Somewhat surprisingly, no such correlation emerged for theta-band Δ(FS-S) (Pearson's correlation coefficient $r = 0.40$, $p = 0.158$). This result suggests that the delta-band neural measure FS-S, which we take as in index of phonetic-feature encoding, is sensitive to increases in the perceived clarity of speech that come with access to prior knowledge. Interestingly, however, for a majority of subjects (11 out of 14), and despite the positive correlation with behaviour, our index of phoneme level processing actually decreased with prior information, a finding that runs counter to our primary hypothesis. This suggests the possibility of a second effect involving a suppression of responses at this hierarchical processing level to the P condition relative to NP (*t*-test on FS-S: $p = 0.003$).

In order to clarify the factors that led to the suppressive effect of prior knowledge on the delta-band cortical index FS-S, the various model performances were compared for the NP and P stimuli. It is important to re-emphasise that each pair of NP and P stimuli had identical physical properties. Therefore significant differences in the corresponding scalp responses must be due to some combination of the following two factors: 1) it could be related to the enhancement of perceived clarity with prior information, a suggestion that is supported by our abovementioned positive correlation between $\Delta_{Clarity}$ and Δ(FS-S), and 2) it could be related to the fact that the P stimulus is a repetition of a previously presented stimulus, while the NP stimulus is always a first presentation. If the latter is a factor in causing a reduction in delta-band EEG prediction accuracy, it should be evident in the pattern of model performances. Indeed, model performances for the NP and P stimuli exhibited different patterns in terms of the relative model performances (**Figure 4.3C**). Specifically, the model performances for NP were similar to those for clear speech, with the combined FS-model performing better than both S and F (ANOVA: $F(1.14,14.87) = 7.22$, $p = 0.014$; post hoc paired *t*-test comparisons of FS with all other models: $p = 0.012$, $p = 0.001$ for S and F respectively). This was not the case for the responses to the P stimuli. In fact FS performed better only than F, while no significant difference emerged when compared with S (ANOVA: $F(1.29,16.72) = 4.24$, $p = 0.04$; post hoc paired *t*-test comparisons of FS with all other models: $p = 1$, $p = 0.001$ for S and F respectively). The model predictions were lower for NP stimuli than for clean speech, but had a similar relative performance pattern (paired *t*-test on S: $p = 0.88$; F: $p = 0.04$; FS: $p = 0.01$), which was not particularly surprising given that noise vocoding reduced the intelligibility of the NP stimuli, but did not make them completely unintelligible.

This pattern of results suggests that the delta-band EEG predictability measures are sensitive to the effect of prior knowledge, and that this prior knowledge primarily affected the interaction between acoustic (S) and phonetic (F) speech models, rather than any individual model performance. In fact, no significant effect (enhancement nor suppression) emerged for any single speech representation/model between NP and P (paired *t*-test on S: $p = 0.16$; F: $p = 0.16$; FS: $p = 0.29$). Unlike in the delta-band, EEG predictability in the theta-band did not exhibit different results patterns for NP and P stimuli. No significant difference emerged between FS and S for either NP or P stimuli, suggesting that cortical entrainment measures in the theta-band are not affected by differences in perceived clarity (NP stimuli: ANOVA, $F(1.17,15.16) = 4.83$, $p = 0.039$; post hoc paired *t*-test comparisons: $\boldsymbol{p = 1}$, $p = 0.002$, $p = 0.208$ for **S vs FS**, F vs FS, and S vs F respectively; P stimuli: ANOVA, $F(1.09,14.22) = 5.97$, $p = 0.026$; post hoc paired *t*-test comparisons: $\boldsymbol{p = 1}$, $p = 4.3*10^{-5}$, $p = 0.292$ for **S vs FS**, F vs FS, and S vs F respectively).

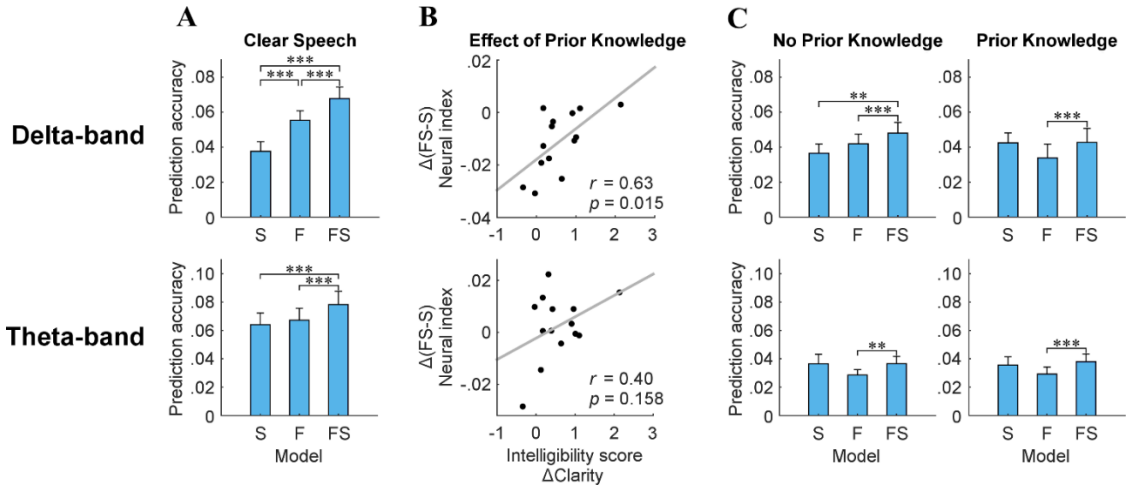


Figure 4.3: The effect of prior knowledge on EEG predictability.
Linear regression was used to fit models known as multivariate temporal response functions between the low-frequency (delta-band and theta-band) EEG and different representations of the speech stimulus. In particular, speech was represented as its spectrogram (S), a time-aligned sequence of categorical phonetic features (F) or a combination of both (FS) (*$p < 0.05$, **$p < 0.01$, ***$p < 0.001$). The difference in performance between the FS- and S-models (i.e., FS-S) is taken as an isolated measure of phoneme-level encoding. **(A)** Correlations (mean ± SEM) between recorded EEG and EEG predicted using the mTRF models for spectrogram (S), phonetic features (F), and their combination (FS) for clear speech. **(B)** A significant positive correlation emerges between the change in perceived intelligibility (measured as Δclarity) and the change in our isolated index of phoneme level delta-band entrainment from NP to P speech segments ($(FS\text{-}S)_P - (FS\text{-}S)_{NP}$) as a result of prior knowledge. **(C)** Correlations (mean ± SEM) between recorded EEG and EEG predicted using the mTRF models for spectrogram (S), phonetic features (F), and their combination (FS) for noise-vocoded speech. In the delta-band, the FS-model performs best for the NP speech segments (no prior knowledge) but not for the P segments (prior knowledge). No significant differences emerge in the theta-band.

4.3.3 Differential effects of prior knowledge on distinct phonetic features

The results so far suggest that prior knowledge affects the EEG-measured cortical tracking of speech and, crucially, the correlation between perceived clarity and FS–S links this effect directly with the cortical processing of phonetic features of speech. To clarify how prior information affects specific speech features, we compared the model-weights across conditions, speech representations, and time-lags in the delta-band (**Figure 4.4**). This analysis was conducted on a broader time-lag window of -100 to 500, which allowed for a clearer contrast between more and less meaningful time-lags. In addition, the TRF-weights shown in the figure were averaged across a set of 12 fronto-central well-predicted electrodes.

The acoustic models, which were fit using the envelope and the 3-band spectrogram of speech, showed stronger average responses in the C condition compared to NP and P, while the weights of NP and P were very similar (**Figure 4.4A**). A more interesting pattern of results emerged for the F-model (**Figure 4.4B**). In particular, there appeared to be differences between the C, P, and NP models in the vowel-based features of the TRF (**Figure 4.4B**). These differences were supported by a simple statistical cluster analysis that compared the phonetic feature TRFs between conditions (uncorrected t-tests at every time-lag and for every feature; **Figure 4.4C**). To examine this in another way, we collapsed the TRFs across phonetic feature categories (Manner of Articulation, Voicing, Vowels, and Place of Articulation) and examined the resulting one-dimensional TRFs across conditions (along with the standard Envelope TRF for comparison; **Figure 4.4D**). A significant suppression of the N1 and P1 components for vowel features emerged for NP and P compared with C (permutation test between NP- and C-models: $p < 0.05$ for -15–85 ms and 195–312 ms; permutation test between P- and C-models: $p < 0.05$ for -15–54 ms and 187–250ms; significant clusters with less than 2 contiguous time-lags were excluded; **Figure 4.4D**). Interestingly, although not significant, the average suppression was greater for P compared to NP. Qualitatively, consonant voicing and place of articulation features resemble the weights for clear speech in the P but not in the NP condition, while no obvious similarity across conditions emerged for manner of articulation features, although there were no statistically significant effects on this.

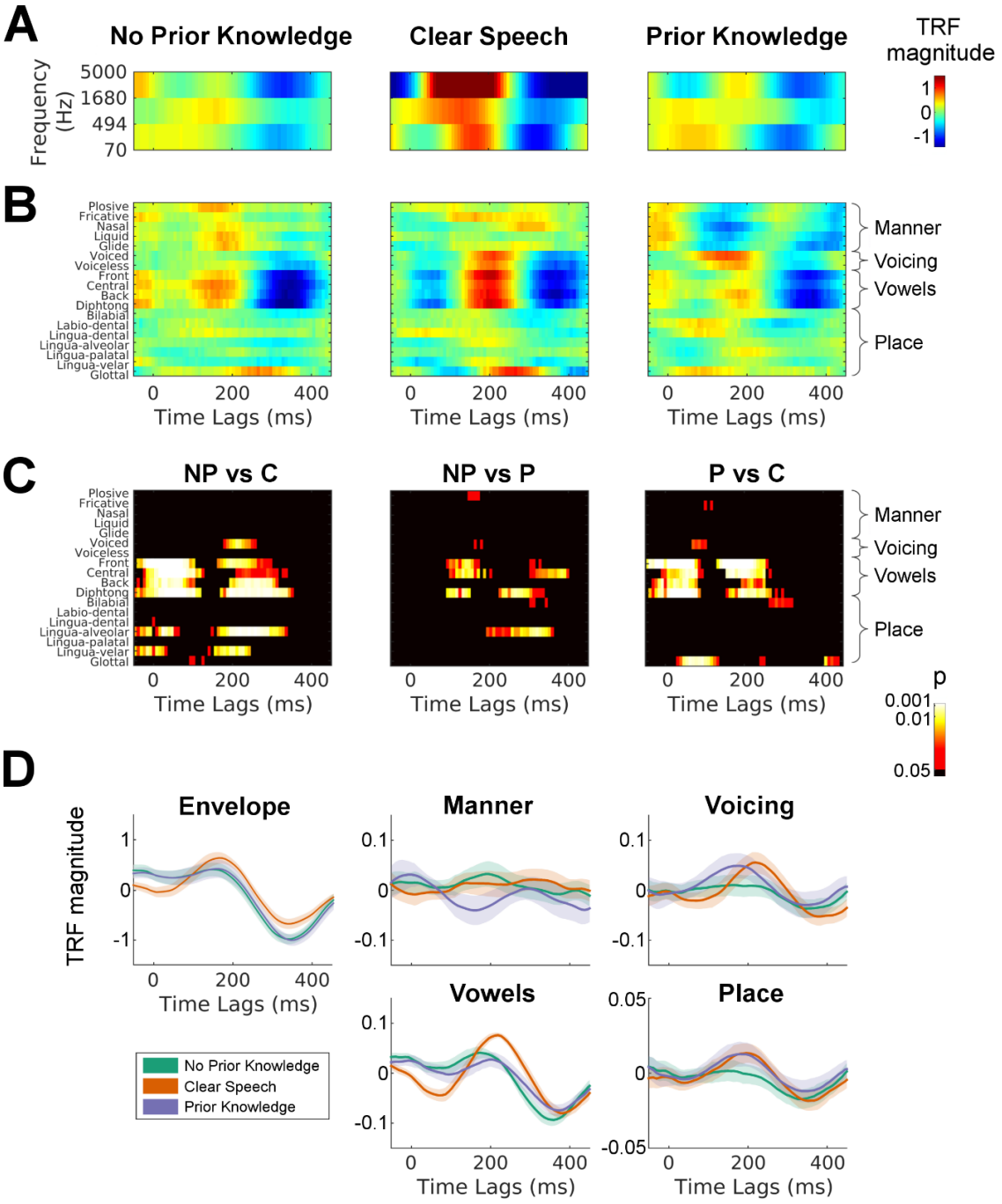


Figure 4.4. The effect of prior knowledge on the temporal response functions.
(A) The TRF (model weights) for the spectrogram representation of speech (S) are shown for all conditions after averaging across 12 selected electrodes (see Section 4.2.4). To allow a direct comparison of all conditions, the TRF for the C-model is shown using only 3 frequency-bands, although the model used in the analysis included all 16 bands. Colours indicate the TRF magnitude (arbitrary units). **(B)** TRF models fit using phonetic features (F) are shown for all conditions. **(C)** F-model weights were compared between each pair of conditions using *t*-tests at each time-lag and phonetic feature. **(D)** To more directly compare the TRF weights between conditions, univariate models are shown for the envelope of speech and for four distinct groups of phonetic features (average weights of each group are reported): manner of articulation, voicing, vowels, and place of articulation.

4.4 Discussion

This chapter investigated the effect of prior knowledge on the cortical entrainment to acoustic and phonetic features of speech using non-invasive EEG and the novel analysis approach introduced in Chapter 3. The results observed for the clear speech reproduced the ones shown in the previous chapter. In the delta-band, a weaker but similar pattern emerged for NP stimuli, which were only partially intelligible because of a severe degradation of their acoustic properties. Crucially, a different result pattern was observed for P stimuli, indicating that prior knowledge modulates the cortical entrainment to speech features. We hypothesised that this phenomenon would be reflected in an increase in the novel measure of cortical entrainment to speech-specific phonetic features that was suggested in Chapter 3 (FS-S). This hypothesis turned out to be partially supported by our data, which exhibited two top-down effects of prior knowledge. The first effect was in line with our hypothesis and took the form of a positive correlation between our neural measure and perceived clarity across subjects. The second ran counter to our hypothesis and took the form of an overall reduction in EEG prediction accuracy for the P stimuli.

4.4.1 A novel isolated index of speech-specific processing

Previous research has failed to find any effect of perceived speech intelligibility on low-frequency cortical tracking of the speech envelope using a perceptual pop-out task (Millman et al., 2015). This is consistent with our findings in that we saw no correlation between perceived clarity and tracking of low-level acoustics (via the S-model). It was only by using differential model performances as our index that we were able to isolate processing at the phonetic-feature level and reveal a relationship. This points to a concern about relying on envelope tracking as a measure of speech processing (Obleser et al., 2012). Specifically, it is highly likely that such a reliance leads to neural indices that reflect multiple, distinct functional processes (Ding and Simon, 2014), making it difficult to determine to what extent the indices reflect speech-specific activity. This might explain why there has been a lack of consistency across studies aimed at examining the effects of speech intelligibility on neural measures of envelope tracking (Howard and Poeppel, 2010; Peelle et al., 2013; Ding et al., 2014). We suggest that our approach represents one way of partially disentangling the multiple processes that must be active during natural speech perception.

The effects of prior knowledge discussed here emerged only in the delta-band of the EEG. This is in line with a current view suggesting that delta- and high-frequency activity (> 40 Hz) are reliable indicators of perceived linguistic representations, while theta-band activity may primarily reflect the analysis of the acoustic features of speech (Ding et al., 2014; Kösem and van Wassenhove, 2016). Indeed one study, in particular, examined the cortical tracking of vocoded speech in background noise and found that delta-band tracking correlated with speech recognition scores across subjects (Ding et al., 2014), a result that corresponds very nicely with our neural-behavioural correlation. However, the specificity of our effects to the delta-band also appears to run counter to other studies examining the relationship between cortical tracking of vocoded speech and intelligibility. Peelle and colleagues (2013) reported significant differences between the cortical tracking of intelligible and unintelligible (vocoded) speech in the theta-band. That said, the authors of that study reported no correlation between their behavioural measures of intelligibility and their theta-band tracking indices. In addition, they did not control for the fact that their intelligibility manipulation (vocoding) covaried with the amount of sensory detail in their stimuli, an issue that we have attempted to address and that has been shown to be important in their more recent work (Blank and Davis, 2016). So it is possible that their theta-band effects actually reflect something other than intelligibility and, therefore, that they do not in fact conflict with our findings.

4.4.2 Neural basis for the counteracting effects of prior knowledge

The emergence of two effects of perceptual pop-out is consistent with previous studies suggesting that prior knowledge may produce counteracting effects (e.g., Tuennerhoff and Noppeney, 2016). One view is that predictions increase the perceived clarity by inducing a better synchronisation of the cortical responses to speech (Peelle et al., 2013), which would produce larger cortical entrainment measures. Along the same lines it has been proposed that increased entrainment measures may reflect the activation of higher-order areas that would have been "inactive" or less responsive when perceived clarity was degraded (George et al., 1999; Davis and Johnsrude, 2003; Peelle et al., 2013; Tuennerhoff and Noppeney, 2016). Both of these ideas are consistent with our positive neural-behavioural correlation across subjects. On the other hand, predictive coding theories assert that prior knowledge of an upcoming stimulus should suppress the measured cortical responses, as those responses are proposed to represent the error

between what is predicted and the bottom-up sensory input (Friston, 2005; Clark, 2013). And this would be consistent with the overall suppression we see in our neural index of phonetic-feature encoding.

In our data, the first effect of prior knowledge on the cortical responses is an enhancement in delta-band tracking that correlates with perceptual clarity. This effect, which emerged from the between-subject analysis, appears to be in line with a recent study from Holdgraf and colleagues (2016) that reported an increase in ECoG activity in auditory cortical areas using a similar pop out paradigm. That study, which could not report on the correlation between such an increase and any behavioral measure of clarity, explained such enhancement as a rapid change in the tuning of cortical responses, which become more responsive to specific speech features when prior information is available. Although directly relating that interpretation to our model-based predictions is not perfectly straightforward, one possible link is through previous work on cross-frequency coupling between high-frequency amplitude and low-frequency phase (e.g. Gross et al., 2013). It may be the case that the increase in high-gamma band amplitude reported by Holdgraf et al. could be linked to the low-frequency tracking enhancements reported in our data. The second effect, which involves an overall suppression of FS–S with prior knowledge, is consistent with the late suppression in left STG shown by Sohoglu et al. (2012) and in line with predictive coding theories. Indeed, because of our experimental design, the stimulus repetition for P trials may contribute to this suppressive phenomenon. On the one hand, it has been hypothesised that such suppressive effects are automatic and due to stimulus-induced neural adaptation (Grill-Spector et al., 2006). On the other hand, the suppression may be a consequence of top-down predictions and could be explained via the theory of predictive coding (Summerfield et al., 2008; Todorovic et al., 2011). However, research on repetition suppression usually involves short, isolated auditory stimuli (e.g. tones), which are very different from the 10-s sentences used in the present study. As such, we are inclined to tentatively suggest that repetition suppression and adaptation will not have played a major role in our findings, but rather that our suppression effects are likely a consequence of predictive coding.

Indeed a review of predictive coding theory has proposed that there may exist two distinct units within our sensory processing hierarchies: representational units and error units (Hohwy, 2013). And this idea fits well with our dual effects. It may be the case that activity from representational units is increased with prior knowledge in our experiment, while activity from error units is suppressed. Future work involving a more balanced

factorial design may be able to more clearly separate these two effects. Such a design should also allow for an analysis of how the cortical tracking measures we have reported here (which we take as an index of the encoding of phonetic features) relate to ongoing gamma- and beta-bands oscillations, given that they have been linked with bottom-up prediction errors and top-down predictions, respectively (Arnal et al., 2011).

4.5 Summary

The present chapter demonstrates that indices of speech-specific cortical processing can be isolated using non-invasive EEG. Specifically, it was assessed that the measure FS-S, which was suggested in the previous chapter, is sensitive to the effect of prior information during the perception of degraded speech. This study suggests that the improvement in perceived intelligibility and the impact of predictability resulting from the prior knowledge have opposite effects on the encoding of phonemes. Firstly, the availability of prior information enhanced the perceived clarity of degraded speech, which was positively correlated with changes in phonetic-feature encoding across subjects. In addition, prior knowledge induced an overall reduction of this cortical measure, which we interpret as resulting from the increase in predictability.

Chapter 5. Causal cortical dynamics of a predictive enhancement of speech intelligibility

5.1 Introduction

The previous chapter demonstrated that prior knowledge modulates EEG measures of cortical entrainment to speech features. This phenomenon emerged only for phonological features (neural index FS-S) and was suggested to reflect the cortical integration of prior information with auditory input. This finding suggests that measures of cortical entrainment can be used to better understand exactly how, where, and when predictive mechanisms influence speech perception, which are issues that are central to the understanding of this cortical network (Norris et al., 2016). However, the previous chapter could not fully account for such complex dynamics because of the low spatial resolution of the experimental data. Here, we describe a pop-out experiment that has the goal of investigating the effects of prior knowledge on the dynamics in temporal and frontal cortical areas using source-space MEG signals.

As we know from Chapters 2 and 4, the cortical hierarchy that underpins speech comprehension is characterised by both bottom-up and top-down signals (Hickok and Poeppel, 2007; Peelle et al., 2010; Gross et al., 2013; Bornkessel-Schlesewsky et al., 2015). In particular, top-down connections may constitute a neural basis for the integration of prior knowledge in the speech processing network (Davis and Johnsrude, 2007; Wild et al., 2012; Lewis and Bastiaansen, 2015). Although recent studies provided important insights into the functions of distinct key areas of the speech network, a number of fundamental questions remain unanswered, especially regarding the temporal dynamics and interactions between cortical sites. Previous studies either did not have the

temporal resolution (Blank and Davis, 2016), the cortical coverage (Holdgraf et al., 2016) or, as in Chapter 4, the spatial resolution to characterise precise dynamics between regions in the speech network. Studies with the requisite spatiotemporal resolution (Sohoglu et al., 2012; Sohoglu and Davis, 2016), focused on cortical (de)activation, rather than indexing the representational content that may underlie such responses, i.e., the neural encoding of speech features.

Here, we seek a better understanding of the spatiotemporal cortical dynamics that underpin the integration of prior knowledge and sensory input by using measures of MEG power, neural encoding of the speech envelope, and a connectivity analysis in key cortical areas. To this end, MEG data from a perceptual "pop-out" experiment (Millman et al., 2015) were re-analysed. In Millman et al. (2015) perceptual pop-out (e.g. Davis et al., 2005) was used to change the percept of physically identical tone-carrier vocoded speech sentences (in short, tone-vocoded sentences) from unintelligible to intelligible during data acquisition. The pop-out effect was obtained by preceding the presentation of some of the vocoded sentences with the original, unprocessed version of the stimulus.

As in Chapter 4, the pop-out approach dissociates the effects of (top-down) prior knowledge from (bottom-up) changes in sensory information (Sohoglu et al., 2012; Millman et al., 2015; Blank and Davis, 2016; Holdgraf et al., 2016; Sohoglu and Davis, 2016). Crucially, the availability of anatomical MRI scans for every subject enabled analyses in source-space (rather than in sensor-space). In particular, bespoke MEG beamformer-based analysis was used to estimate neural sources in bilateral locations of interest (Millman et al., 2015), corresponding to HG, STS, MTG, and IFG. These regions have been shown to provide distinct contributions to the speech recognition process and to represent progressively higher levels of the speech perception hierarchy (Davis and Johnsrude, 2003; Scott and Johnsrude, 2003; Hickok and Poeppel, 2007; Peelle et al., 2010; Peelle et al., 2013; Mesgarani et al., 2014; Overath et al., 2015; Leonard et al., 2016; Sohoglu and Davis, 2016; Tuennerhoff and Noppeney, 2016). Similarly to Sohoglu et al. (2012), it was hypothesised that an early activation of IFG would precede cortical processing in temporal areas such as STG and STS. Furthermore, the high spatiotemporal resolution of this data may be sufficient to reveal both top-down and bottom-up dynamics. In this context, this study seeks to clarify how fast such top-down and bottom-up modulation occur and which specific areas (and order) they involve.

The neural encoding of speech was estimated using measures of cortical entrainment to the temporal envelope of speech sentences (Lalor et al., 2009; Crosse et

al., 2016b). As the functional roles and interpretations of the cortical entrainment phenomenon are still debated (Ding and Simon, 2014), the first goal of this study was to determine whether and how entrainment to the speech envelope is affected by perceptual pop-out and could therefore entail sensitivity to the integration of prior knowledge with sensory information and the consequent change in perceived intelligibility. Secondly, we aimed to investigate the top-down/bottom-up dynamics of the pop-out effect by using measures of cortical entrainment, event-related power, and effective connectivity.

The findings described in this chapter were presented at the Society for Neuroscience 56th annual meeting as a poster (November 2016) and have been submitted as a research article as: "Causal cortical dynamics of a predictive enhancement of speech intelligibility", *in review*.

5.2 Methods

The present study is based on re-analyses of a previously published MEG study on perceptual pop-out (Millman et al., 2015).

5.2.1 Participants

Sixteen right-handed native English speakers (10 males; mean age = 29.2 years ± 7.8 years, age range = 20-48 years) took part in this experiment. The participants reported normal hearing and no history of neurological disorders.

5.2.2 Speech stimuli

Short-duration sentences spoken by an adult British English male (BKB/IHR corpus; e.g. Macleod and Summerfield, 1987; Foster et al., 1993) were used as the speech stimuli. The duration of each speech sentence was approximately 1.5-s, and the duration of each epoch was extended to 2.5-s through the addition of approximately 1-s of silence to the end of each sentence. Stimuli were delivered diotically to participants via Etymotic insert earphones (Etymotic Research ER30, Elk Grove Village, IL) at a comfortable sound level.

The sentences were "The kettle boiled quickly" (always unintelligible), and "The floor was quite slippery" (the pop-out sentence), and "She ironed her skirt" (always intelligible). The intelligibility of all three speech sentences was degraded by using a tone-carrier vocoder (e.g. Dudley, 1939). A tone-vocoder with only three carriers was used to produce vocoded stimuli that were unintelligible prior to exposure to the original,

unprocessed version of the same sentence. The carrier frequencies were 225, 1047, and 4861 Hz. The temporal envelopes at the output of each channel were extracted using half-wave rectification and smoothing. The cut-off frequency of the low-pass filter used to smooth the extracted temporal envelope varied depending on the carrier frequency. Specifically, the cut-off frequency was set to half the equivalent rectangular bandwidth (e.g. Moore and Glasberg, 1983) of each channel (24, 68, and 274 Hz for each of the carrier frequencies, respectively). The temporal envelopes extracted from each band were then summed to form the broadband speech temporal envelope for each sentence.

Technical note: Tone-vocoding versus Noise-vocoding

This experiment used tone-vocoding for speech degradation, while Chapter 4 used noise-vocoding. The main difference between the two approaches lies in the carrier signal: In tone-vocoding it is a sine-wave with a frequency equal to the centre frequency of the vocoder channel; Noise-vocoding uses a noise carrier with a centre frequency and bandwidth equal to the vocoder channel. Another similar method, called sine-wave speech, is instead based on a formant tracking approach rather than a pre-defined frequency-band separation as in vocoding approaches.

5.2.3 Experimental paradigm

The experiment was carried out in a single session for each subject and was composed of three parts, as outlined in **Figure 5.1**. The main rationale was to present participants with unintelligible vocoded speech, which was perceived as unintelligible regardless of how many times it was repeated. Crucially, prior exposure to the original clear version of vocoded speech enhances perceived intelligibility when listening to the vocoded version (i.e., the "pop out" effect; Davis et al., 2005). In this experiment, participants were presented with repetitions of the original version of the *intelligible* stimulus at the beginning of the MEG session. During *block 1*, MEG data were recorded as participants listened to the three vocoded speech sentences and to silent trials (2.5-s duration). These auditory conditions were named *Pop-out*, *Unintelligible*, *Intelligible*, and *Silent*, and they were presented in random order for a total of 100, 100, 50, and 50 times respectively. In block 1, participants could comprehend only the *Intelligible* stimulus. At the end of block 1, participants were presented with repetitions of the original version of the *Pop-out* stimulus during a *training block* with no neural recordings. Finally, MEG recordings were

performed during *block 2*, which was physically the same as *block 1*, but with the crucial difference that both the *Intelligible* and the *Pop-out* stimuli were perceived as intelligible.

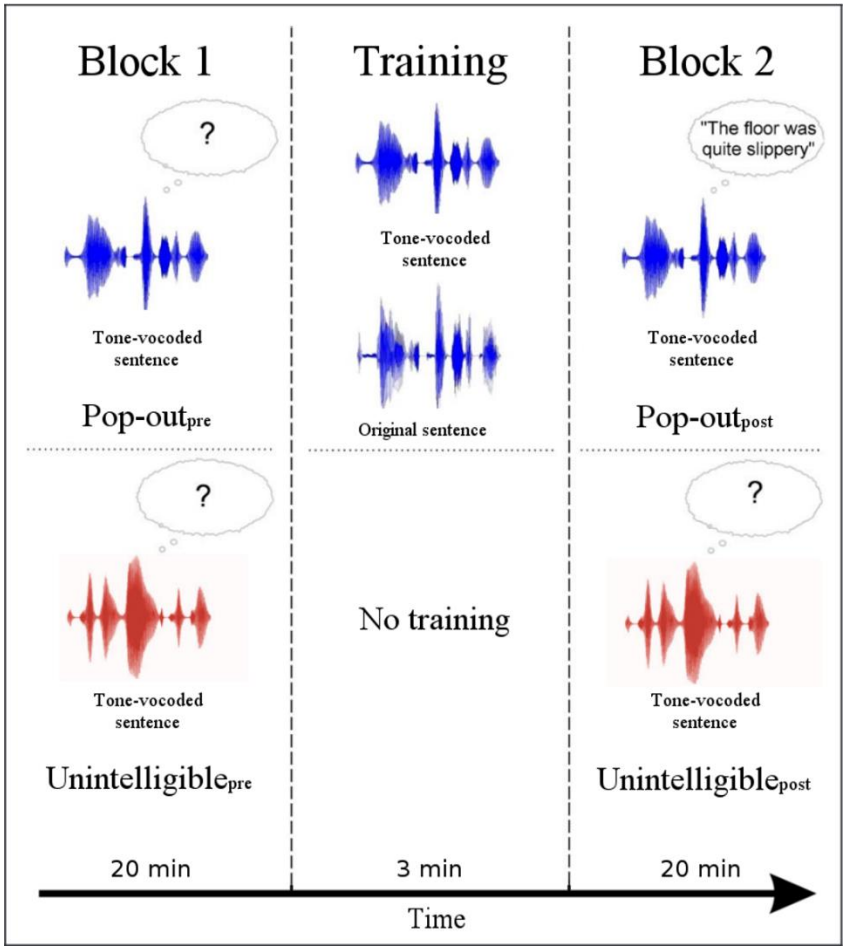


Figure 5.1: A pop-out experiment to isolate predictive perceptual enhancement of speech.
Adapted from Millman et al. (2015). MEG data were recorded while participants listened to speech sentences that were degraded using 3-channel tone-carrier vocoded speech (*Pop-out*$_{pre}$ and *Unintelligible*$_{pre}$ conditions). In block 1 (~20 minutes) both vocoded sentences of interest were perceived as unintelligible. A training block followed (~3 minutes in which MEG data was not recorded) in which participants listened to the vocoded and the original versions of only one of the two sentences. Finally, both vocoded sentences of interest were presented in block 2 (~20 minutes). In the latter case, the pop-out sentence (*Pop-out*$_{post}$ condition) became intelligible after training, whereas the other sentence of interest remained unintelligible (*Unintelligible*$_{post}$ condition) because participants were not exposed to the corresponding original version.

The conditions of interest, i.e. *Unintelligible* and *Pop-out*, recorded before the training block are denoted *Pop-out*$_{pre}$ and *Unintelligible*$_{pre}$. Likewise, the same conditions recorded after the training block are referred to as *Pop-out*$_{post}$ and *Unintelligible*$_{post}$. In addition, probe trials were presented during each block (2.5 seconds duration; 25 per block). During a probe trial, participants were played an auditory cue, which prompted them to respond, using a button box, and indicate a binary intelligibility rating (intelligible or unintelligible) for the last sound they heard. The resulting values were used to verify

that the desired enhancement of perceived clarity from block 1 to block 2 occurred for the *Pop-out* trials but not for the *Unintelligible* trials. As reported in Millman et al. (2015), due to a technical issue, intelligibility ratings were only recorded for 15 of the 16 participants.

5.2.4 MEG recordings

Data were collected at the University of York, UK, using a Magnes 3600 whole-head 248-channel magnetometer (formerly 4-D Neuroimaging, Inc., San Diego, CA). The signals were recorded at a sample rate of 678.17 Hz and were low-pass filtered online with a cut-off frequency of 200 Hz.

Before recording, individual facial and scalp landmarks (left and right preauricular points, Cz, nasion, and inion) were spatially coregistered using a Polhemus Fastrak System. The landmark locations in relation to the sensor positions were derived on the basis of a precise localisation signal provided by five spatially distributed head coils with a fixed spatial relation to the landmarks. These head coils provided a measurement of a participant head movement at the beginning and end of each data acquisition block.

The raw data in each epoch were inspected visually. Epochs contaminated with either physiological or non-physiological artifacts were manually removed.

5.2.5 Coregistration

For the source-space analyses, the landmark locations were matched with the individual participants' anatomical magnetic resonance (MR) scans using a surface-matching technique adapted from Kozinska et al. (2001). T1-weighted MR images were acquired with a GE 3.0-T Signa Excite HDx system (General Electric, Milwaukee, WI) using an eight-channel head coil and a 3-D Fast Spoiled Gradient Recall sequence: repetition time/echo time/flip angle = 8.03 ms/3.07 ms/20°, spatial resolution of 1.13 mm × 1.13 mm × 1.0 mm, in-plane resolution of 256 × 256 × 176 contiguous slices. The individuals' data were spatially normalised to the Montreal Neurological Institute (MNI) standard brain, based on the average of 152 individual T1-weighted structural images (Evans et al., 1993). The source-space grid for each participant was initially defined in MNI space and linearly transformed back to individual MRIs.

5.2.6 Beamformer-based analyses

An MEG beamformer estimates the contribution of a given grid point in the brain to the signal measured at the MEG sensors. Independent beamformers (spatial filters) are constructed for each grid point. Each beamformer is an optimal spatial filter dedicated to a given grid point. The outputs of these spatial filters are often termed "virtual electrodes".

In this study, a vectorised, linearly constrained minimum-variance beamformer (Van Veen et al., 1997; Huang et al., 2004) was used to obtain the spatial filters with a multiple-spheres head model (Huang et al., 1999). The beamformer grid size was 5 mm. The three orthogonal spatial filters were implemented as a single 3-D system (see Johnson et al., 2011). In this beamformer framework, the orientation of each spatial filter is a crucial free parameter that should be specified for the metric of interest. In this study, an independent, unsupervised approach was chosen for the optimisation of the spatial-filter orientation. In particular, a principal component analysis (PCA) was performed to extract one dominant signal from a space of 163 orientations with equal spatial distribution. This choice allowed for all analyses (power, envelope entrainment, and causality) to be performed using the same source-space data and avoided possible overfitting due to the dependency between spatial-filter orientation and entrainment analysis parameters.

5.2.7 Locations of interest

The aim of this study was to characterise the effects of prior knowledge (and of the consequent enhancement in speech intelligibility) on the activity in several bilateral key locations in the speech comprehension hierarchy and their interactions (e.g. Hickok and Poeppel, 2007): These key locations included, as depicted in **Figure 5.2**, HG, STS, [MNI: ±61, -22, 0] (coordinates taken from Overath et al., 2015); posterior MTG, [MNI: ±55, -46, -4] (coordinates taken from Lau et al., 2008); and IFG, [MNI: ±54, 18, 20] (coordinates taken from Sohoglu et al., 2012). As in the work of Millman and colleagues (Millman et al., 2013; Millman et al., 2015), left and right HG were manually seeded because the anatomy of HG varies considerably among individuals (e.g. Rademacher et al., 2001).

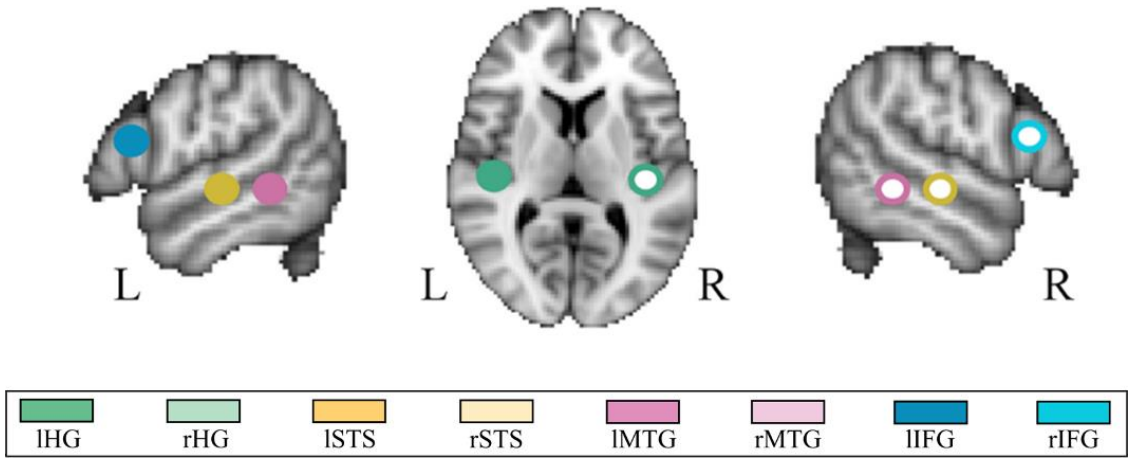


Figure 5.2: Schematic of the cortical locations of interest.
A standard MNI brain is used to display the 4 bilateral cortical areas selected for this study.

5.2.8 Frequency bands of interest

Spatial filters from the LOIs were generated using a time window of 2000 ms, including 500 ms prior to stimulus presentation. Broadband (1–45 Hz) data from the conditions of interest (Pop-out, Unintelligible) were projected through the spatial filters in the first instance so that all analyses (i.e., power envelope, entrainment and causality) could be carried out using the same spatial filters orientation. Contributions from more specific brain rhythms were assessed by spectrally filtering the broadband source-space signal in the frequency bands delta (1–4 Hz), theta (4–8 Hz), alpha (8–15 Hz), beta (15–30 Hz), and gamma (30–45 Hz) using Chebyshev Type 2 digital filters in both a forwards and backwards direction to remove phase-distortion.

5.2.9 Event-related power analyses

Event-related fields time-locked to stimulus onset were derived for *Pop-out*$_{pre}$, *Pop-out*$_{post}$, *Unintelligible*$_{pre}$, and *Unintelligible*$_{post}$. Whilst fMRI studies are limited to overall measures of cortical activity over relatively long time windows, the current analysis also investigated the temporal dynamics of the cortical responses to speech. This information is conveyed by means of the cumulative event-related power, where cumulative power at time t is calculated as the sum of the squares for the time window [0, t]. The use of this measure allowed for a clearer visualisation of overall trends across the duration of a sentence compared to the more common point-by-point power analysis. Baseline correction was applied using the pre-stimulus time interval from -0.5 to -0.2 ms.

After calculating the cumulative power measures for each individual condition, a combined measure was derived for each location and frequency band using the same

contrast as for the normalised difference used by Millman et al. (2015): [(*Pop-out*$_{post}$ – *Pop-out*$_{pre}$) – (*Unintelligible*$_{post}$ – *Unintelligible*$_{pre}$)]. The reasoning behind the use of this specific contrast is that *Unintelligible*$_{post}$ – *Unintelligible*$_{pre}$ accounts for MEG differences due to repetition, while *Pop-out*$_{post}$ – *Pop-out*$_{pre}$ also reveals the effects of prior knowledge and its effects on speech intelligibility. This overall measure reflects the source-space MEG power enhancement due to the perceptual pop-out. On a similar note, the same contrast could be re-written as [(*Pop-out*$_{post}$ – *Unintelligible*$_{post}$) – (*Pop-out*$_{pre}$ – *Unintelligible*$_{pre}$)], where *Pop-out*$_{pre}$ – *Unintelligible*$_{pre}$ accounts for MEG variation due to low-level physical differences between the two vocoded sentences, while *Pop-out*$_{post}$ – *Unintelligible*$_{post}$ also reflects the effects of prior knowledge and its effects on speech intelligibility.

5.2.10 Cortical entrainment analyses

The mapping between stimulus and cortical activity was estimated using the mTRF system identification approach (see Section 2.3.5). In particular, differently from Chapter 4, the procedure involved identifying a mapping from source-space MEG signal to the speech envelope (*backward-modelling*) that optimised the following linear model:

$$\hat{s}_{loc}(t) = \sum_{\tau=\tau_i}^{\tau_i+winSize} r(t+\tau, loc)\, g(\tau, loc), \qquad (5.1)$$

where $\hat{s}_{loc}(t)$ is the estimated speech envelope using the MEG signal from a location of interest *loc*, $r(t+\tau, loc)$ is the MEG response time lag τ_i and location *loc*, and $g(\tau, loc)$ is the linear decoder for the corresponding time lag and location. The objective was to reconstruct the underlying speech envelope and to compare the quality of such reconstructions across experimental conditions and cortical locations of interest. The decoder $g(\tau, loc)$ was optimised for each condition using leave-one-out cross-validation while maximising the correlation between $\hat{s}_{loc}(t)$ and $s(t)$ (Crosse et al., 2016b; mTRF Toolbox: http://sourceforge.net/projects/aespa/). At the cross-validation step *i*, data from all trials but trial$_i$ were used to fit a model and to reconstruct an estimate of the envelope for trial$_i$. The procedure was then repeated for all trials, which allowed us to derive such an estimate for every trial.

In order to control for overfitting, we conducted a parameter search to select the optimal value for the regularisation parameter λ, i.e., the value of λ that produces the highest mean correlation between envelope and its estimates across all trials. This mean

correlation, which was measured by calculating Pearson's correlation (*r*), represents a quantitative measure of cortical entrainment to the envelope of speech.

A window of time-lags with a size of 200 ms (Crosse et al., 2016b) was shifted from shorter to longer latencies with steps of 50 ms (from -50 – 150 ms to 200 – 400 ms), which allowed investigation of the spatiotemporal dynamics of the effect of interest. The quality of fit was estimated for each window. Note that models are fit on the whole duration of a sentence (1.5 s), while the time-lag window-size refers to how many data-points of the MEG signal are used to reconstruct one single point of the speech envelope.

Here, the contrast used for the power analysis [(*Pop-out*$_{post}$ – *Unintelligible*$_{post}$) – (*Pop-out*$_{pre}$ – *Unintelligible*$_{pre}$)] was decomposed into *Pop-out*$_{post}$ – *Unintelligible*$_{post}$ and *Pop-out*$_{pre}$ – *Unintelligible*$_{pre}$. Only the *Pop-out*$_{post}$ – *Unintelligible*$_{post}$ contrast involves the perceptual pop-out effect, while the *Pop-out*$_{pre}$ – *Unintelligible*$_{pre}$ contrast may only reflect differences in cortical entrainment due to low-level physical differences between the two vocoded sentences. The choice of these contrast measures ensures that the resulting effects are neither due to repetition suppression nor perceptual learning. In fact such factors could emerge here only when the number of prior presentations of the stimuli differ between conditions, i.e., for contrasts between different blocks. The chosen approach is valid for the entrainment analysis as it involves subtraction of prediction correlations. However, these measures could not be used in the event-related power analysis while preserving the time-domain, as the subtraction of MEG power would be dominated by low-level physical differences between the two sentences.

5.2.11 Network effective connectivity analysis

Brain connectivity measures are used to infer neuronal spatiotemporal interactions which index and predict task-relevant changes in cognitive states and behaviour. Whilst methods such as dynamic causal modelling (DCM) require a set of possible hypotheses for the neurobiological system of interest (Stephan et al., 2007), there exist approaches that do not impose such a constraint and relies on data-driven analyses (Granger, 1969; Ding et al., 2006). Here, an exploratory dynamical framework for neuronal system identification was used to assess the effect of prior information on the *effective connectivity* between cortical areas of interest in temporal and frontal lobes (*effective connectivity* denotes asymmetric causal dependences between brain regions). To this end, the source

information flow toolbox (SIFT, Delorme et al., 2011) was used to investigate such causal effects in the context of bottom-up/top-down cortical information flow.

SIFT was used to fit Vieira-Morf models on the source-space MEG data. A direct directed transfer function (dDTF; Korzeniewska et al., 2003) measure was used to estimate the direct causality between pairs of cortical locations of interest ('directed' indicates that nodes are connected by directional edges; 'direct' clarifies that direct information flow is isolated from indirect/spurious edges). dDTF can be interpreted as frequency-domain conditional granger causality (GC) measure and is effective in removing spurious indirect causal influence between brain sources (Kus et al., 2004). For each connection, this analysis looks for significant causal interactions in the time-frequency domain. The SIFT toolbox was provided with the broad-band source-space data (1 – 45 Hz). Data preprocessing consisted of constant detrending, and time and ensemble normalisation (Ding et al., 2000) with the following model parameters: model-order = 18, window-size = 300 ms, step-size = 30 ms. Furthermore, SIFT performs a frequency analysis by means of a segmentation-based linear vector autoregressive model (similar to a short-time Fourier transform) (Ding et al., 2000). Model validation was performed by checking for its stability and the whiteness of the residuals by means of the autocorrelation function (ACF) test (Lütkepohl, 2007; Delorme et al., 2011). The smallest model order that led to stability and whiteness for all experimental conditions was selected.

As for the event-related power analysis, the contrasts between ($Pop\text{-}out_{post}$ – $Pop\text{-}out_{pre}$) and ($Unintelligible_{post}$ – $Unintelligible_{pre}$) were used to investigate the effect of prior knowledge in the spatio-spectral MEG domain. Because the time domain was involved, the connectivity contrasts $Pop\text{-}out_{post}$ – $Pop\text{-}out_{pre}$ and $Unintelligible_{post}$ – $Unintelligible_{pre}$ were derived first (**Figure 5.7**), as they involved the same physical stimulus and, therefore, the time dimension could be preserved. By collapsing the dDTF measure along the time domain, it was possible to compare the results for *Pop-out* and *Unintelligible* stimuli.

5.2.12 Statistical analysis

All statistical analyses were conducted using Wilcoxon signed-rank tests (paired if possible), except where otherwise stated. All numerical values are reported as mean ± SD. In the cortical power and in the entrainment analyses, permutation-based cluster

statistics (N = 1000 repetitions) were used to correct for multiple comparisons while keeping in consideration that results for neighbouring time points or time windows are not independent. In the connectivity analysis, Bonferroni correction was applied by taking into consideration both the number of frequency bins and the number of nodes.

5.3 Results

5.3.1 Behavioural intelligibility ratings

The responses made during the probe trials were analysed to confirm that the *Pop-out* sentences were perceived as more intelligible in block 2, i.e., after exposure to the unprocessed speech. The low intelligibility ratings for the *Pop-out*$_{pre}$ (mean = 15.7%, SD = 34%), *Unintelligible*$_{pre}$ (mean = 11.1%, SD = 17.3%), and *Unintelligible*$_{post}$ (mean = 17.8%, SD = 28.9%) sentences indicate that they were perceived as unintelligible. The intelligibility ratings for *Pop-out*$_{post}$ (mean = 93.5%, SD = 15.2%) were significantly greater than the ones for *Pop-out*$_{pre}$ (Wilcoxon signed rank test, $N = 15$ subjects, $p = 0.001$) and *Unintelligible*$_{post}$ (Wilcoxon signed rank test, $N = 15$ subjects, $p = 0.001$), indicating that prior knowledge induced perceptual pop-out in the *Pop-out* condition only. In the *Unintelligible* control condition, the ratings for *Unintelligible*$_{pre}$ were not significantly different than those for *Unintelligible*$_{post}$, (Wilcoxon signed rank test, $N = 15$ subjects, $p = 0.15$), indicating that the *Unintelligible* condition was an adequate control for temporal order effects. In block 1, there were no significant differences in the rating for *Pop-out*$_{pre}$ and Unintelligible$_{pre}$ (Wilcoxon signed rank test, $N = 15$ subjects, $p = 0.95$), indicating that both conditions were perceived as similarly unintelligible.

5.3.2 Distinct effects of perceptual pop-out on neurophysiological power

Event-related power enhancement showed significant effects for STS, MTG, and IFG (Wilcoxon signed rank test, N = 16 subjects, $p < 0.05$; cluster statistics were used to correct for multiple comparisons for all the tests in this section). **Figure 5.3** shows how this measure varies across the whole sentence duration (1.5 s) across all locations and frequency bands of interest. Sustained delta-band power enhancement was measured in left STS ($p < 0.05$) and left MTG ($p < 0.05$). MTG showed significant left lateralisation of such enhancement (paired Wilcoxon signed-rank test, N = 16, $p < 0.05$). Importantly, these sustained effects did not emerge for other frequency bands. A different pattern of

results was measured in IFG, which showed early (~100-550 ms) left-lateralised effects in gamma-band ($p < 0.05$; significant left lateralisation, paired Wilcoxon signed-rank test, N = 16, $p < 0.05$) and right-biased enhancements for longer latencies for broadband power (~600-1300 ms respectively; $p < 0.05$; significant right lateralisation emerged for the broadband signal: paired Wilcoxon signed-rank test, N = 16, $p < 0.05$).

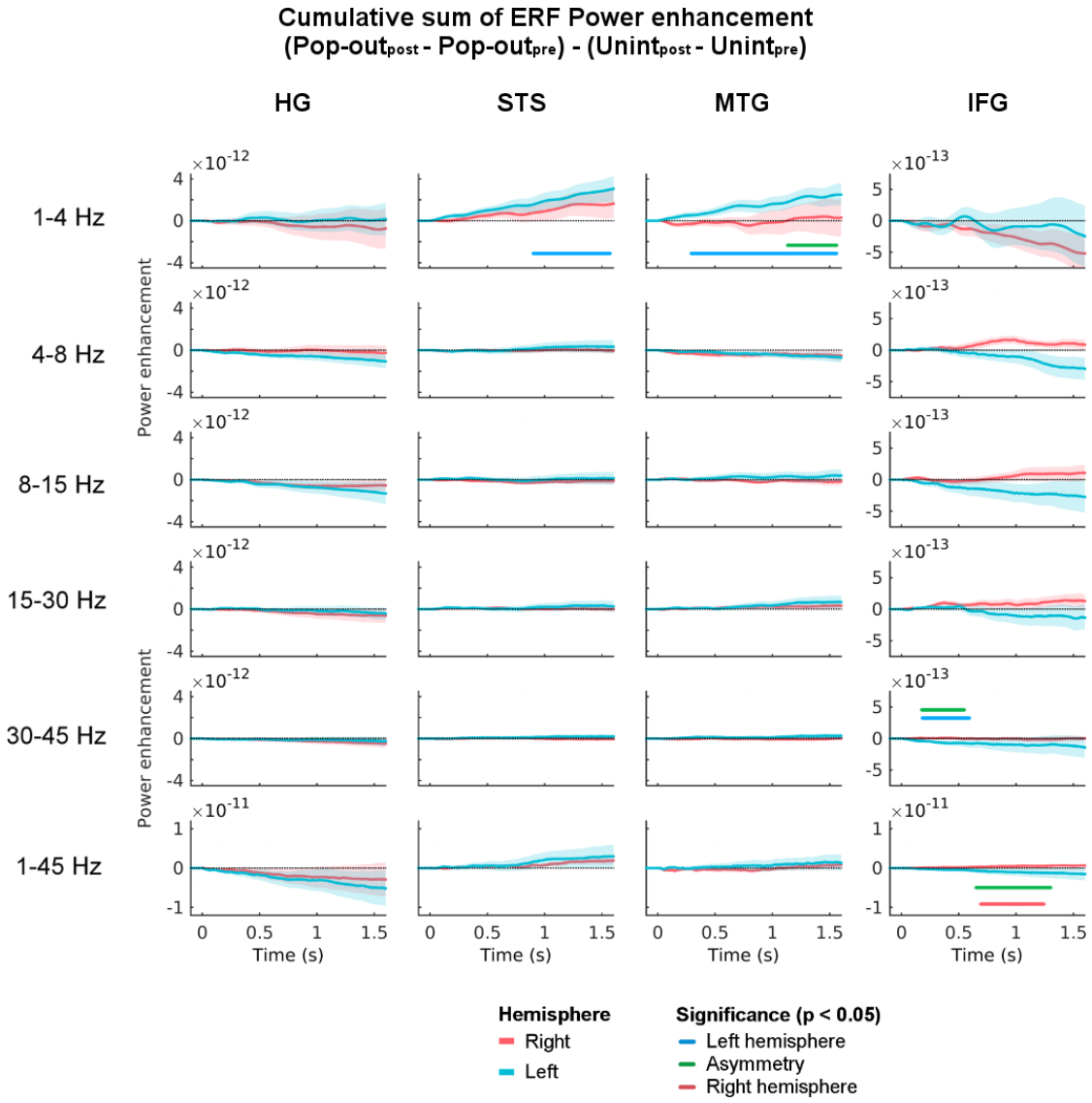


Figure 5.3: Perceptual pop-out determines changes in source-space MEG power.
Event-related fields time-locked to sentence onsets were derived and power measures were calculated for each time sample from a latency of zero. Baseline correction was applied using the time interval from -0.5 to -0.2 ms. The cumulative difference in power [(*Pop-out*$_{post}$ – *Pop-out*$_{pre}$) – (*Unintelligible*$_{post}$ – *Unintelligible*$_{pre}$)] is reported here for all frequency bands and cortical sources of interest (left and right hemispheres are directly compared to investigate possible asymmetries; values on the y-axis are reported in arbitrary units). Significant power change and hemispheric asymmetries are marked with horizontal lines (Wilcoxon signed rank tests, $N = 16$, $p < 0.05$; cluster statistics across the time dimension was used to correct for multiple comparisons). Note that IFG and results for broad-band signals are shown on different ordinate scales as they exhibited event-related power at different magnitude than the other locations and frequency bands of interest.

This pattern of enhancement and suppression of cortical activity, as depicted in **Figure 5.3**, is based on the normalised difference measure [(*Pop-out*$_{post}$ – *Pop-out*$_{pre}$) – (*Unintelligible*$_{post}$ – *Unintelligible*$_{pre}$)], which isolates the effects of perceptual pop-out (Millman et al., 2015). Therefore, the present experiment successfully elicited an increase in perceived intelligibility that was reflected in the source-space MEG signal. In particular, the sustained effects in left STS and left MTG suggest that the perceptual change in speech intelligibility is underpinned by enhanced activity in those cortical locations (Peelle et al., 2013; Blank and Davis, 2016; Tuennerhoff and Noppeney, 2016). Shorter suppressive effects also emerged in HG and IFG. Although this result indicates that the MEG data reflects some effects of prior knowledge, this specific measure was insufficient to assess top-down/bottom-up cortical interactions and, importantly, it did not show overwhelming suppressive effects due to prior knowledge that would have supported other studies based on similar pop-out paradigms (Sohoglu et al., 2012; Blank and Davis, 2016; Sohoglu and Davis, 2016). One reason may lie in the event-related approach itself, which is ill-suited for the relatively long sentences used in this experiment (Crosse et al., 2016b): The fast dynamics of conversational speech hamper the investigation of long latency responses, as they overlap with the early responses to subsequent sounds. This is not the case when short isolated stimuli are used such as single words (Sohoglu et al., 2012; Sohoglu and Davis, 2016). For this reason, additional analyses were used that aimed at eliciting more targeted indices of cortical activity with the goal of determining the precise spatiotemporal dynamics due to the effect of prior knowledge.

5.3.3 Prior knowledge modulates top-down and bottom-up envelope entrainment

Cortical entrainment was used to determine how accurately the broadband envelope of speech could be reconstructed from the source-space MEG signals of individual participants, as measure by correlation (Pearson's *r*). A change in this correlation when *Unintelligible*$_{post}$ was compared with *Pop-out*$_{post}$ (block 2) was used to quantify changes in cortical entrainment due to perceptual pop-out (**Figure 5.4**). A significant entrainment enhancement involving almost all the cortical areas of interest emerged for the delta-band (1–4 Hz) at different latencies. Specifically, an early sustained enhancement of left cortical areas arose, starting from consecutive time windows for, in this order, IFG, HG, STS, and MTG, from -50–150 ms, 0–200 ms, 50–250 ms, 100–300 ms respectively

(Wilcoxon signed-rank test, $N = 16$, $p < 0.05$; correction for multiple comparisons was performed using a cluster statistics that takes into account dependencies across the time dimension). This result suggests that delta-band activity in left IFG may initiate the early propagation of envelope entrainment within the left-hemisphere in a bottom-up direction from HG, to STS, and finally to MTG. The enhancement in such areas was sustained until the lag-window 150–350 ms for all left cortical areas, while only left MTG also showed enhancement for 200–400 ms. Cortical enhancement in right cortical areas emerged only in the window 100–300 ms for IFG, and in the windows from 150 ms for STS and MTG, which were excluded after the correction for multiple comparisons performed using cluster statistics. Envelope entrainment in the theta-band (4–8 Hz) showed an early suppression with perceptual enhancement in all cortical areas of interest, while reaching (non-corrected) significance in left STS, right HG, STS, and IFG. However, only the suppressive effect in IFG remained significant after application of the cluster correction.

The link between these results and the perceptual pop-out effect was controlled by performing the same analysis on envelope reconstruction correlations derived for the sentences of interest in block 1. Any differences in the reconstruction accuracies between $Unintelligible_{pre}$ and $Pop\text{-}out_{pre}$ (**Figure 5.5**) represent a baseline that accounts for physical differences between the two vocoded sentences that do not involve the pop-out effect. The effects seen for block 2 did not emerge for stimuli presented in block 1 and, importantly, no significant increase in entrainment emerged for any time-lag window. The interaction between block and entrainment enhancement was formally tested by means of an three-way ANOVA analysis that, while considering the three factors *block*, *cortical location*, and *time-lag window*, found a significant main effect of *block* in both delta-band ($F(1,1280) = 166.50$; $p = 6.71*e^{-36}$) and theta-band ($F(1,1280) = 78.29$; $p = 2.90*e^{-18}$).

These findings provide detailed information on the effects of prior knowledge on the early cortical dynamics underlying continuous speech processing. In particular, they indicate that the availability of higher-level information in the upcoming stimulus, which enhances the perceived intelligibility of the speech sentences, increasing the early delta-band tracking of the envelope of speech in left IFG. Enhanced delta-band entrainment was also measured in other left cortical areas at progressively longer latencies, suggesting that information flow initiated by left IFG then propagates to primary auditory cortex, followed by superior temporal and posterior middle temporal areas within the left hemisphere.

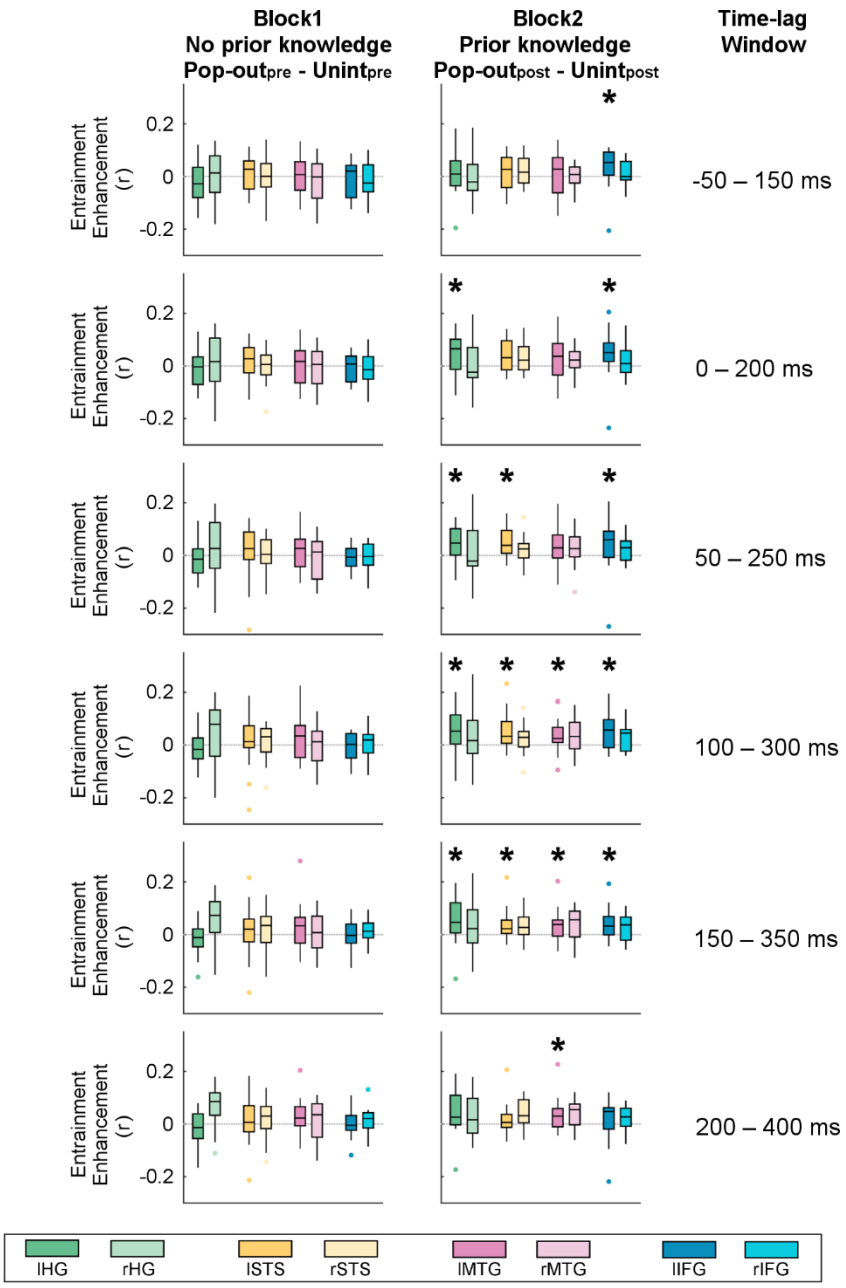


Figure 5.4: Prior information induces top-down dynamics of delta-band entrainment to the speech envelope.
TRFs were evaluated from the cortical responses to the envelope of the stimulus (backward modeling approach; window size: 200ms). The quality of fit was estimated by calculating correlations between the speech envelope and its reconstructions using cross-validation. Differences between *Pop-out* and *Unintelligible* sentences were calculated for both blocks 1 and 2 (before and after exposure to the original sentence in the *Pop-out* condition). This procedure was repeated for each cortical location of interest. Significant effects emerged for delta-band (Wilcoxon signed-rank test, N = 16, *$p < 0.05$; cluster statistics across the time dimension was used to correct for multiple comparisons). **(A)** Differences between *Pop-out* and *Unintelligible* stimuli for block 2, where only the *Pop-out* stimulus was perceived as intelligible. Significant effects represent cortical entrainment enhancement induced by prior information and physical differences between *Pop-out* and *Unintelligible* sentences. **(B)** Differences between *Pop-out* and *Unintelligible* sentences for block 1, where both stimuli were unintelligible. Significant contrasts are caused by differences in entrainment brought about by physical differences between the stimuli.

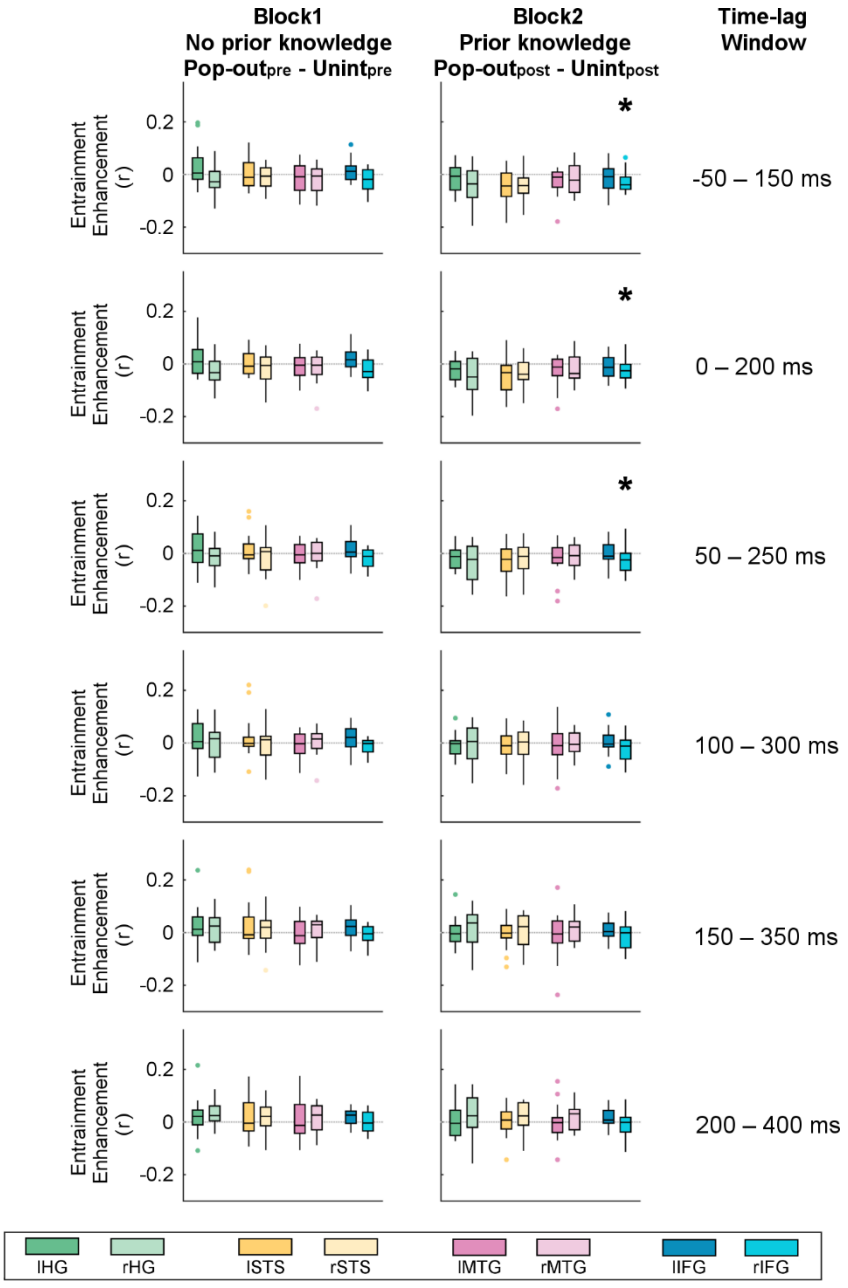


Figure 5.5: Effects of prior information on the theta-band entrainment to the speech envelope. TRFs were evaluated from the cortical responses to the envelope of the stimulus (backward modelling approach; window size: 200ms). The quality of fit was estimated by calculating correlations between the speech envelope and its reconstructions using cross-validation. Differences between *Pop-out* and *Unintelligible* sentences were calculated for both blocks 1 and 2 (before and after exposure to the original sentence in the *Pop-out* condition). This procedure was repeated for each cortical location of interest. Some significant effects were identified for theta-bands (Wilcoxon signed-rank test, N = 16, *p < 0.05; cluster statistics across the time dimension was used to correct for multiple comparisons).

5.3.4 Effects of prior knowledge on cortico-cortical dynamics

The cortical entrainment analysis provided insights on how perceptual pop-out affects the encoding of the speech sentences measured with MEG. However, this approach is constrained to the speech features chosen for the analysis, which in this case consists of

the speech envelope. Furthermore, while both the entrainment and power analyses study the effects of perceived intelligibility in each individual cortical area separately from the others, there are approaches that allow the explicit investigation of the causal interaction between cortical areas. Importantly, these approaches enable the study of how the pop-out effect modifies the spatiotemporal dynamics of the speech comprehension network, without biasing the analysis to specific features of speech (e.g., speech envelope). Therefore, further analysis was conducted with the goal of obtaining complementary insights on the cortical mechanisms of integration of prior information during speech comprehension.

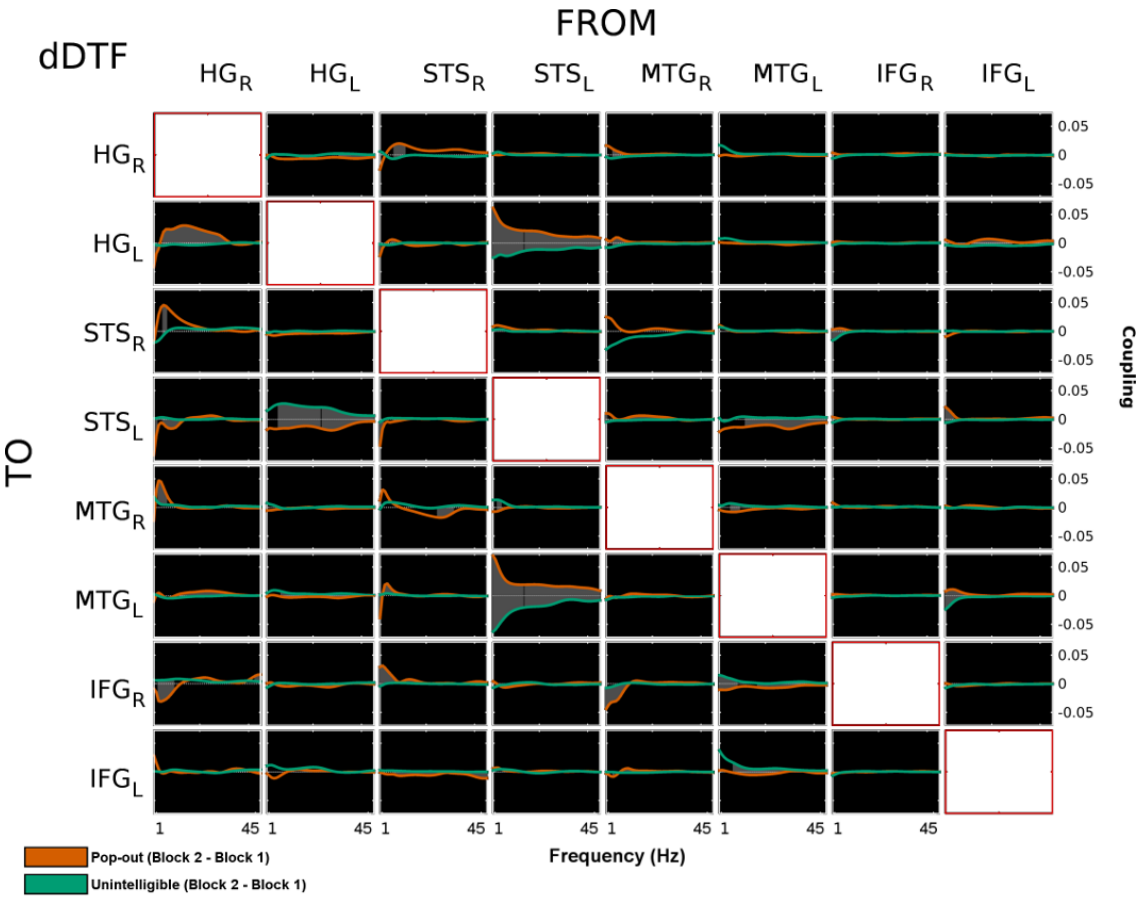


Figure 5.6: A source-space connectivity analysis: Low-level top-down and higher-level bottom-up connections are enhanced when prior information is available.
Frequency grid showing the event-related grand-average (N = 16) change in brain dynamics induced by prior information across frequency, and cortical locations (SIFT; Delorme et al., 2011). dDTF (Korzeniewska et al., 2003) were calculated for an 8 node model, including all locations of interest. These can be interpreted as frequency-domain conditional granger causality measures. Frequency grids show the dDTF contrast from block 1 to block 2, for *Unintelligible* (green) and *Pop-out* (orange) stimuli. Significant differences between stimulus type are highlighted with the grey shaded area (paired Wilcoxon signed rank test, N = 16, $p < 0.05$; Bonferroni correction was applied).

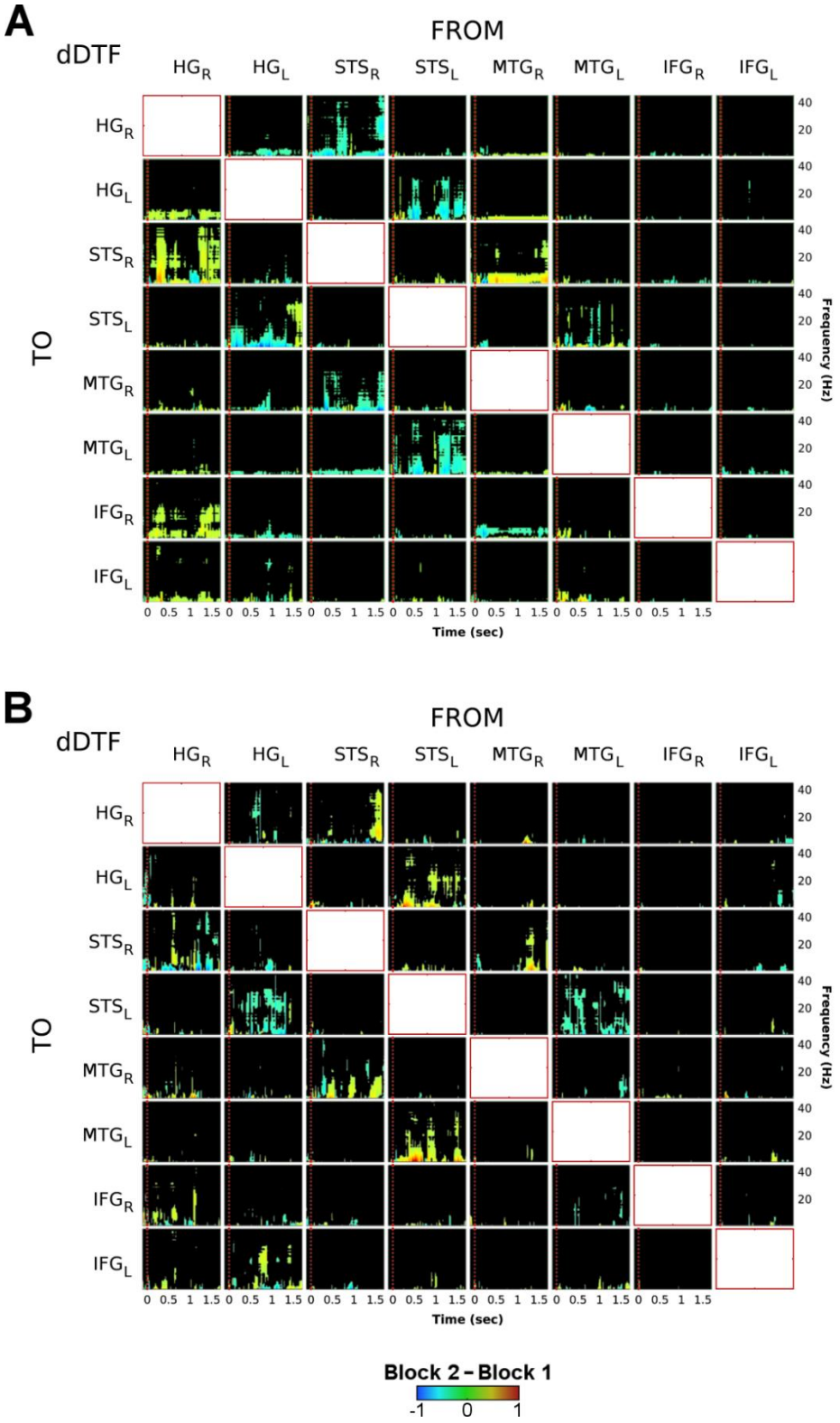


Figure 5.7: Source-space connectivity analyses for distinct speech stimuli.
Time-frequency grid showing the event-related grand-average (N = 16) change in brain dynamics induced by prior knowledge across time, frequency, and cortical locations (SIFT; Delorme et al., 2011). Direct Directed Transfer Function measures (dDTF; Korzeniewska et al., 2003) were calculated for an 8-node model, including all locations of interest. These can be interpreted as frequency-domain conditional granger causality measures. **(A, B)** Time-frequency grids showing the normalised dDTF contrast from block 1 to block 2, for **A**) *Unintelligible* and **B**) *Pop-out* stimuli. The baseline is -100 to -10 ms and the vertical red dashed lines indicates latency of the beginning of a sentence (t = 0 ms). The dDTF contrast measures were thresholded for statistical significance using the 95th percentile of all measured values.

Here, this was achieved by estimating effective connectivity measures within the eight nodes network of locations of interest (bilateral HG, STS, MTG, and IFG). The space-frequency grid in **Figure 5.6** shows significant changes in top-down and bottom-up connectivity between cortical locations of interest due to perceptual pop-out. The strongest effects of prior information were an increase of the top-down link between left STS-left HG, a stronger bottom-up connection for left STS-left MTG, and a stronger inter-hemispheric connection from left HG to right HG. The connectivity between left HG-left STS exhibited both an increase of top-down and decrease of bottom-up information flow over almost the whole frequency band of interest. In contrast, the link between left STS-left MTG showed an enhanced bottom-up information flow due to prior knowledge, with a concomitant suppression of top-down connection. Finally, the significant links to and from IFG involved mainly low frequencies (< 15 Hz). **Figure 5.7** shows this same result with the inclusion of the time dimension for both *Unintelligible* and *Pop-out* stimuli (**Figure 5.7A and B** respectively).

5.4 Discussion

The cortical mechanisms that underpin the integration of prior knowledge with sensory input during continuous speech comprehension are poorly understood. This chapter demonstrated that non-invasive MEG measures are sensitive to the predictive effects of prior knowledge on perceived speech intelligibility. Furthermore, we provided insight into the cortical spatiotemporal dynamics of this network that have implications for current views of the cortical underpinnings of speech comprehension.

5.4.1 Low-frequency envelope entrainment reflects perceived speech intelligibility

Measures of cortical entrainment to speech features, in particular to the speech envelope, are powerful tools to investigate the cortical mechanisms of continuous speech processing (Ahissar et al., 2001; Aiken and Picton, 2008; Nourski et al., 2009; O'Sullivan et al., 2014; Crosse et al., 2016a). However, it remains unclear to what extent envelope entrainment reflects the encoding of sensory information that is specific to speech (Ding and Simon, 2014; Di Liberto et al., 2015; Zoefel and VanRullen, 2016) and how entrainment is affected by higher-order processes (such as the integration of prior knowledge; Holdgraf et al., 2016). It has been argued that speech intelligibility impacts on the phase of ongoing

neural oscillation in (left) temporal cortex, which was interpreted as suggesting that linguistic information affects neural oscillations (Peelle et al., 2013; Park et al., 2015). Yet, those studies did not disentangle the effects of perceptual from physical (acoustic) differences, as the intelligibility was modulated using physical manipulations of the speech stimuli (e.g., noise-vocoding, time-reversing).

A previous analysis of the same MEG dataset used in the present study did not reveal any significant effect of prior knowledge on phase-locking to the speech envelope (Millman et al., 2015), quantified using theta-band coherence and cross-correlation measures between stimulus envelope and source-space neural signals. In particular, the latter approach is sub-optimal when studying responses to stimuli with speech-like statistics (Crosse et al., 2016b) and did not allow an exploration of the temporal dynamics and specificity of the pop-out effect. Chapter 4 aimed at relating ongoing EEG to particular features of a speech stimulus and found a positive correlation between delta-band entrainment to phoneme-level features and perceived speech intelligibility using a pop-out paradigm similar to the one employed here. However, no significant effect of prior knowledge emerged specifically for the envelope entrainment. It may be the case that the effect of prior knowledge on cortical entrainment is so subtle – relative to the entrainment to the acoustic energy of the stimulus itself – that the use of an imaging modality with higher spatial resolution than sensor-space EEG, combined with a stronger focus on time-domain analyses, is required to reveal it.

In contrast to the previous work from Millman et al. (2015), the results shown in **Figure 5.4** clearly indicate that low-frequency envelope entrainment is affected by perceptual pop-out. Importantly, this effect was related specifically to the increase in perceived speech intelligibility and not to an effect of stimulus repetition or perceptual learning (Sohoglu and Davis, 2016). Furthermore, the result was not a consequence of physical differences between the two stimuli as no effects on envelope entrainment were observed for block 1, in which both stimuli (*Pop-out*, *Unintelligible*) were unintelligible. Indeed, this finding is in line with the notion that cortical entrainment to the envelope of speech is prominent in the delta- and theta-bands (Ahissar et al., 2001; Aiken and Picton, 2008; Giraud and Poeppel, 2012; Gross et al., 2013; Keitel et al., 2017). Specifically, our results agree with a recent view suggesting that delta-band entrainment is strongly linked with speech intelligibility (Ding and Simon, 2013; Ding et al., 2014; Ding and Simon, 2014) and with the formation of temporal predictions (Arnal et al., 2015). On the other hand, theta-band entrainment has been associated with the acoustic properties of the

speech envelope (Ding and Simon, 2013; Peelle et al., 2013). However, our results suggest a role of theta-band entrainment in the predictive mechanisms that support an increase in the perceived intelligibility of speech.

5.4.2 Top-down and bottom-up effects of prior knowledge during speech perception

Speech comprehension is thought to involve the active integration of prior knowledge with sensory input through predictive top-down cortical mechanisms (Davis and Johnsrude, 2007; Wild et al., 2012; Blank and Davis, 2016; Leonard et al., 2016; Tuennerhoff and Noppeney, 2016), however the exact spatiotemporal dynamics of this process remain unclear. In particular, there is strong evidence that such effects occur in a broad network of cortical areas, including regions in temporal cortex and frontal areas such as IFG (Sohoglu et al., 2012; Park et al., 2015; Sohoglu and Davis, 2016). Our results advance the current understanding of this cortical network by providing new insights on its precise spatiotemporal dynamics.

Here, we show that delta-band cortical entrainment to the envelope of speech is affected by prior knowledge, and that this effect rapidly propagates in a top-down manner, from left IFG to left HG (time windows [-50, 150] ms and [0, 200] ms respectively), and only subsequently reaches STS and then MTG (**Figure 5.4**). Interestingly, such a top-down phenomenon did not emerge for theta-band (**Figure 5.5**), which instead showed early suppression of entrainment. This finding supports the notion that neural oscillations at different rates may contribute in distinct ways to predictive mechanisms in speech comprehension (Gross et al., 2013; Ding and Simon, 2014; Fontolan et al., 2014; Kösem and van Wassenhove, 2016). In line with previous studies on speech comprehension, the early cortical entrainment measured in IFG suggests the involvement of a broader network that includes temporal and frontal sites (Hickok and Poeppel, 2007; Obleser and Kotz, 2009; Park et al., 2015; Sohoglu and Davis, 2016). Additionally, our results demonstrate that IFG exerts influence on regions in temporal cortex, supporting speech intelligibility through entrainment to the envelope of speech.

The cortical entrainment analysis indicated that the effects of prior knowledge entailed a rapid top-down propagation that affected all left cortical areas of interest between the time windows [-50, 150] ms and [100, 300] ms. This result suggests that the effects of prior knowledge occur on a much shorter temporal scale than previously reported by similar pop-out studies (Sohoglu et al., 2012; Sohoglu and Davis, 2016). One

explanation may be found in the use of short sentences in the present experiment. In contrast to isolated words, such stimuli are suitable for deriving entrainment measures in addition to event-related responses. Moreover, meaningful sentences enable the study of cortical responses to linguistic features. For these reasons, the present work highlights the importance of investigating both the cortical responses to isolated speech units and to continuous speech stimuli.

The cortical entrainment analysis provided also a detailed picture of the top-down/bottom-up effects of prior knowledge in the speech comprehension network. However, the quantitative measures were biased toward a specific feature of speech: the temporal envelope. Furthermore, the order in which cortical entrainment emerges in different areas does not imply causality. This motivated further analysis of the MEG responses to unveil causal dynamics to which linear mapping between stimulus and neural recording would be insensitive. A connectivity analysis was performed on the source-space MEG signals to determine effective (directional) direct causality between the cortical regions of interest (**Figure 5.6**). It is important to clarify that our estimate of causality is constrained by two factors: 1) It is specific to dDTF measures, which can be thought as in the same domain as Granger causality and 2) it is limited to the 8 cortical sites of interest. This analysis showed that perceptual pop-out increases the top-down (and suppresses the bottom-up) information flow between left HG and left STS. This is in line with theories of predictive coding which would explain such effects as an early top-down modulation, which increases the readiness of HG to process the upcoming stimulus, and the subsequent reduction of the bottom-up prediction error (Friston and Kiebel, 2009; Clark, 2013). A second effect was the enhancement of the bottom-up causal link from left STS to left posterior MTG. The use of short meaningful sentence stimuli (rather than isolated words) may be crucial for the interpretation of this result; such stimuli enable processing at the syntactic and semantic levels. In this context, previous research suggested the emergence of both activity suppression, at hierarchical levels where speech information is predictable, and activity enhancement, at levels that were previously not (or less) active (George et al., 1999; Tuennerhoff and Noppeney, 2016). Our results suggest that posterior MTG is strongly involved in the processing of higher-level features of speech related to intelligibility (Lau et al., 2008; Turken and Dronkers, 2011; Henseler et al., 2014; Zhang et al., 2015; Tuennerhoff and Noppeney, 2016). The focus on local interactions was essential to reveal the various pattern of this “vertical” information flow.

Finally, our results suggest the involvement of a broader network of cortical areas than previously modelled (Sohoglu and Davis, 2016; Tuennerhoff and Noppeney, 2016) for the predictive effect of prior information, including left frontal and bilateral temporal lobes. Our findings suggest that characterisation of the fast cortical dynamics in response to continuous speech requires neuroimaging technologies with sufficient spatial coverage and high temporal resolution. In line with previous studies, this finding suggests that early integration of prior knowledge in IFG enhances the readiness of the network to the expected incoming stimulus in lower-order temporal areas (Friston, 2005; Sohoglu and Davis, 2016). We contend that frontal areas constitute a crucial part of such a cortical network (c.f. Tuennerhoff and Noppeney, 2016).

5.5 Summary

In summary, this chapter provided detailed spatiotemporal evidence of the top-down and bottom-up dynamics of prior knowledge on speech intelligibility using a pop-out paradigm. First, our results indicate that non-invasive measures of envelope entrainment are sensitive to the predictive effect of prior knowledge. Second, prior knowledge induces rapid information flow in delta-band cortical signals that initiates in left IFG and subsequently bottom-up propagates to HG, STS, and MTG in the left hemisphere. This indicates that left IFG may induce rapid (< 50 ms) top-down modulation of lower levels of the speech processing hierarchy. In line with the notion of predictive coding, an effective connectivity analysis revealed that perception of intelligible speech sentences increased the top-down information flow from left STS to left HG and enhanced the bottom-up causal link from left STS to left MTG, suggesting the involvement of left posterior MTG in the processing of intelligible speech features.

Chapter 6. Indexing cortical entrainment to natural speech at the phonemic level: Methodological considerations for applied research

6.1 Introduction

A significant number of people worldwide suffer from some form of speech and language impairment. These can arise as a consequence of developmental disorders (Leonard, 2014) or from a decline in related cortical functions (e.g., through ageing, psychosis, injury; Ross et al., 2007; Kemper and Anagnopoulos, 2008; Mesulam et al., 2014). A better understanding of the underlying speech processing network and the ability to identify a specific impairment within that network are crucial to developing clinically useful assessments of speech and language in these populations.

Speech and language impairment can disrupt one's ability to understand auditory speech and efficiently communicate in a number of ways, which correspond to different symptoms. In this context, standardised assessment of such impairments is usually pursued using a number of behavioral tests (e.g., non-verbal hearing, speech, and language tests; standardised test of intelligence, Tomblin et al., 1996; Ford and Dahinten, 2005; Gardner et al., 2006). However, these measures are inadequate at capturing the full extent of a person's impairment and should be considered only as one aspect of a comprehensive assessment process (Flanagan et al., 1997; Mody and Belliveau, 2013). Furthermore, some of these measures cannot be derived for some groups such as infants or participants with reading impairment or no reading skills.

A complementary approach is to "directly" investigate the causes that underpin such conditions, rather than evaluate "indirect" effects on specific behavioral markers. In this sense, neuroimaging provides an opportunity to derive measures directly related to the cortical processing of speech in the human brain. In particular, noninvasive, safe, functional brain measurements (EEG, MEG, fMRI, NIRS) have now been proven feasible for use with both children (starting at birth) and adults (Aslin and Mehler, 2005; Kuhl et al., 2005; McNealy et al., 2006; Kuhl, 2010). In this context, the framework defined in Chapter 3 provides a potential novel methodology for investigating natural speech encoding under a variety of conditions and in a variety of cohorts. This could include research on the causes of speech impairments in particular cohorts by deriving direct indices of cortical activity at specific levels of the speech processing hierarchy using non-invasive EEG. However, short experimental times are preferable in applied research (Mirkovic et al., 2015), whereas the methodology introduced in Chapter 3 used a recording time of 72 minutes per subject, which may constitute an obstacle when studying particular cohorts (e.g., young children).

The present chapter introduces an extension of this framework that allows for a significant reduction of the experimental time needed to derive such indices of phoneme-level cortical entrainment. While not many electrodes are needed for this approach to be effective (by construction; see forward modelling approach in Crosse et al., 2016b), the ability to use the framework with small amounts of data from individual subjects is uncertain. To clarify this issue, an extensive analysis was conducted to assess the minimum experimental time needed to detect meaningful cortical responses. The goal of the analysis was to show that it is possible to utilise short data sets across multiple subjects to make inferences about speech processing in individual subjects. Specifically, we aimed to show that we can robustly index phoneme-level processing in the context of natural speech in cases of limited amounts of experimental data. The results of this study were published as a journal article as: "Indexing cortical entrainment to natural speech at the phonemic level: Methodological considerations for applied research", *Hearing Research*, February 2017.

6.2 Material and methods

The present study is based on re-analyses of the dataset collected for Chapter 3. Please refer to Section 3.2 for information on the participants and the EEG experimental paradigm.

6.2.1 Data Preprocessing

EEG data were digitally filtered between 1 and 8 Hz using a Chebyshev Type 2 zero-phase filter. In order to reduce the processing time, all EEG data were then down-sampled to 64 Hz. EEG channels with a variance that exceeded three times that of the surrounding channels were labelled as bad channels, and replaced by an estimate calculated using spherical spline interpolation (EEGLAB; Delorme and Makeig, 2004). All channels were then re-referenced to the average of the two mastoid channels with the goal of maximising the EEG responses to the auditory stimuli (Luck, 2005).

6.2.2 Speech Representations

Following Chapter 3, mTRFs were estimated based on four distinct representations of the speech stimulus:

1. Broadband amplitude envelope (E);
2. Spectrogram (S);
3. Phonetic features (F);
4. Finally, we propose a model that combines F and S (FS).

Please refer to Section 3.2.5 for a detailed description of these speech representations.

6.2.3 Model Evaluation

As performed in Chapter 3, a model-based analysis was conducted to quantify how well the EEG reflects the encoding of the various speech representations. The idea is to fit a model (i.e., an mTRF) that describes the forward mapping from a speech representation to the EEG and then to test that model by seeing how accurately it can predict EEG from a new trial. This is quantified using a correlation coefficient (Pearson's *r*) and the procedure is then repeated for each trial (leave-one-out cross-validation). A single prediction correlation value was then derived by averaging these correlations over all

trials and a set of electrodes of interest (see Section 3.2.7). Note that silent time intervals were removed from the correlation evaluation (the same intervals were removed from all speech representations).

For each participant, predictions of their EEG signals were derived using mTRFs that were fit on data of that specific subject (*subject-specific models*). This approach, which was used also in Chapter 3, was compared to a subject-independent method. This approach consisted of using models obtained by averaging the subject-specific mTRFs obtained from all other subjects (*generic models*). In other words, models fit on 9 subjects were averaged and then used to predict the EEG of the left-out subject.

The mTRF approach requires three key considerations when carrying out model-based predictions of EEG data: 1) the channels to be predicted; 2) the choice of time-lags between stimulus and data to optimise prediction; and 3) the choice of the regularisation parameter λ. The optimisation procedures performed in Chapter 3 were also used here.

6.2.4 Multi-Dimensional Scaling analysis

The mTRF mapping functions produced by the above analysis can be informative about how particular speech features are represented in the EEG. One useful technique for analysing these multivariate mapping functions is known as multidimensional scaling (MDS). The same procedure performed in Chapter 3 was also used here.

6.2.5 Statistical Analyses

Unless otherwise stated, all statistical analyses were performed using a repeated measure, one way ANOVA to compare distributions of Pearson correlation values across models and to compare *F*-Score classifications across response intervals. The values reported use the convention *F*(*df, dferror*). Greenhouse-Geisser corrections were made if Mauchly's test of sphericity was not met. All post-hoc model comparisons were performed using Bonferroni corrected paired *t*-tests.

6.3 Results

6.3.1 Neural evidence for phonetic processing in generic models

To investigate whether phoneme level cortical activity can be indexed using a generic modelling approach, mTRFs (Crosse et al., 2016b) were built to describe the linear mapping from each speech representation to the EEG scalp-recorded signal. The quality of these predictions, measured with Pearson's correlation, indexed how well the EEG reflects the processing of low- and high-level speech features. **Figure 6.1A,B** show this result when the EEG predictions were derived using a subject-specific model, i.e., given a subject, the predictions of their EEG signal were obtained using a model fit on that same subject using cross-validation across trials. While no significant difference emerged between the S- and F-models, the model fit on the combination of the two (FS) produced the highest EEG prediction correlations (ANOVA: $F(3.0,7.0) = 12.1$, $p = 0.004$; post-hoc paired *t*-test comparisons of FS with all other models: $p = 0.001$, $p = 0.005$, $p = 0.023$ for E, S, F respectively). As previously argued in Chapter 3, this result indicates that low-frequency EEG indexes the cortical entrainment to categorical phoneme-level features of speech.

A similar analysis was conducted to assess whether the same effect emerged when using the generic modelling approach. In particular, the same processing steps as in the subject-specific case were performed with one important difference: when predicting the EEG for a given subject, we used models that were fit to data from all the other subjects. Because EEG responses vary across subjects as a result of cortical folding, EEG prediction correlations for generic models were expected to be lower than in the subject-specific approach. This was confirmed by the results in **Figure 6.1C,D** (two-way ANOVA; effect of modelling approach: $F(1,72) = 6.1$, $p = 0.016$). However, crucially, the combined model FS still produced the best EEG predictions in this case (ANOVA: $F(3.0,7.0) = 21.9$, $p = 0.001$; post-hoc paired *t*-test comparisons of FS with all other models: $p < 0.001$, $p < 0.001$, $p = 0.024$ for E, S, F respectively). Again, this suggests that this modelling approach is sensitive to the effects of categorical phoneme-level processing, even when using generic models.

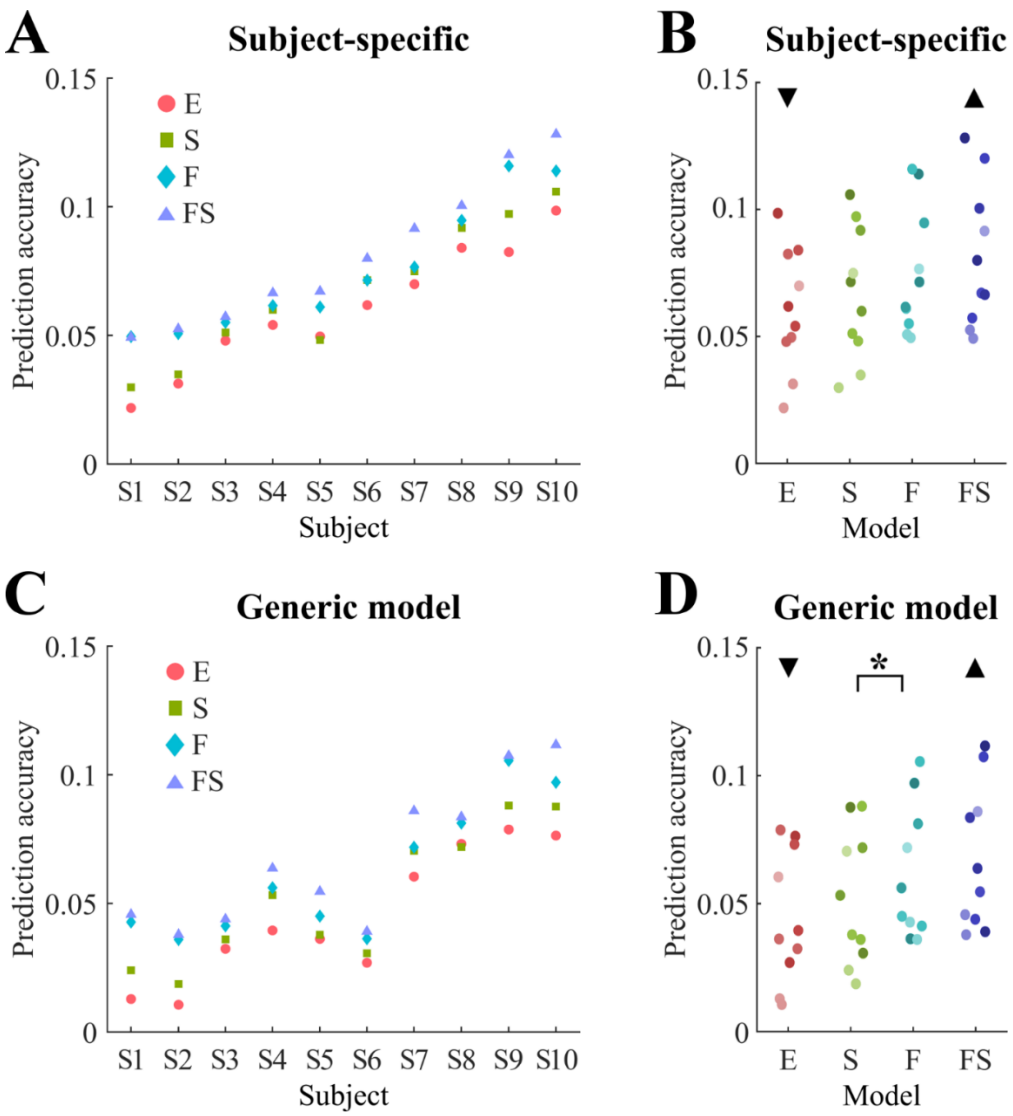


Figure 6.1: Sensitivity of generic models to the encoding of speech features.
EEG data were recorded while subjects listened to natural speech from an audiobook. Speech was represented using E and S (speech acoustics), F (phonetics), and FS (which combines acoustics and phonetics). mTRF were built to describe the mapping from each representation of speech to the EEG recording and used to predict the EEG signal with cross-validation (▲ greater than all others, $p < 0.01$; ▼ smaller than all others, $p < 0.01$; * $p < 0.05$). **(A)** Correlations between EEG and its predictions are shown for each subject and each of the 4 speech representations. The predictions were obtained using speech-specific models, i.e., trained and tested within each subject using cross-validation (subjects were re-arranged according the performance of the FS-model for visualisation purpose). This panel corresponds to Figure 3.1; it is not identical because of minor changes in the data preprocessing (e.g., down-sampling rate). **(B)** The same data is here shown grouped by the 4 speech representations. Each data point refers to a specific subject (a specific colour saturation was assigned to each subject). **(C)** Correlations between EEG and its predictions using a generic model, i.e., trained on all subjects with the exception of the test subject. The subject arrangement is consistent with (A). **(D)** The same values obtained for generic models are here grouped by speech representation. Each data point refers to a specific subject and their colours match the ones shown in (B). Because prediction correlations were calculated using 72 minutes of data, chance level here is effectively zero.

6.3.2 Generic models index phonetic processing for limited experimental time

In Chapters 3 and 4, we suggested that one could potentially derive an isolated measure of phoneme-level processing from the EEG prediction framework. In particular, subtracting the Pearson's correlations for the FS- and S-models (FS-S) should represent such an index of phoneme-level processing. Such a measure would be unsuited for clinical application if it required long experimental times. For this reason, we wished to investigate its robustness as a function of EEG recording time in both subject-specific and generic modelling approaches.

The subject-specific FS-S measure showed a logarithmically increasing prediction accuracy as a function of testing time. Crucially, with less than 30 minutes of EEG data, the FS-S measure was not statistically significantly different from zero across subjects (Wilcoxon signed rank test, $p > 0.05$; **Figure 6.2A**). A significant FS-S measure emerged only when the amount of recording data was greater than or equal to 30 minutes (Wilcoxon signed rank test, $p < 0.05$). Crucially, this issue, which hampers the clinical applicability of such an approach, did not apply to the generic models. The overarching rationale is that the use of data from many subjects produces a model that converges to an effective fit even for short recording times. For instance, since each model was averaged across 9 subjects, an experimental time of 10 minutes corresponded to 90 minutes of data in total, which is more data than was available for any subject-specific model. These considerations are confirmed by the results in **Figure 6.2A**, which shows the significant advantage of the generic modelling approach over the subject-specific approach when up to 40 minutes of data were available from each subject (paired Wilcoxon signed rank test, $p = 0.001$, $p = 0.020$, $p = 0.014$, $p = 0.037$, $p = 0.064$, $p = 0.084$, $p = 0.064$, respectively from 10 to 70 minutes of recording data).

Figure 6.2B provides further support to these considerations by showing how recording time impacts on the mTRF model fit for a particular subject (S4). Qualitatively, the regression weights for the generic models required only 20-30 minutes of data to converge to a stable fit for both S and F. In contrast, the corresponding subject-specific models showed a more gradual and prolonged convergence. Importantly, the subject-specific and generic mTRFs converged to qualitatively similar results (please note that the x-axis was up-sampled and smoothed for visualisation purpose). This observation is confirmed by the numerical result in **Figure 6.1A,C**; in fact the EEG prediction accuracies obtained for subject-specific and generic models for subject 4 (S4) were

comparable. Intuitively, a generic model that does not resemble the corresponding subject-specific model may lead to poor EEG prediction accuracies for the generic model (for example, see S6).

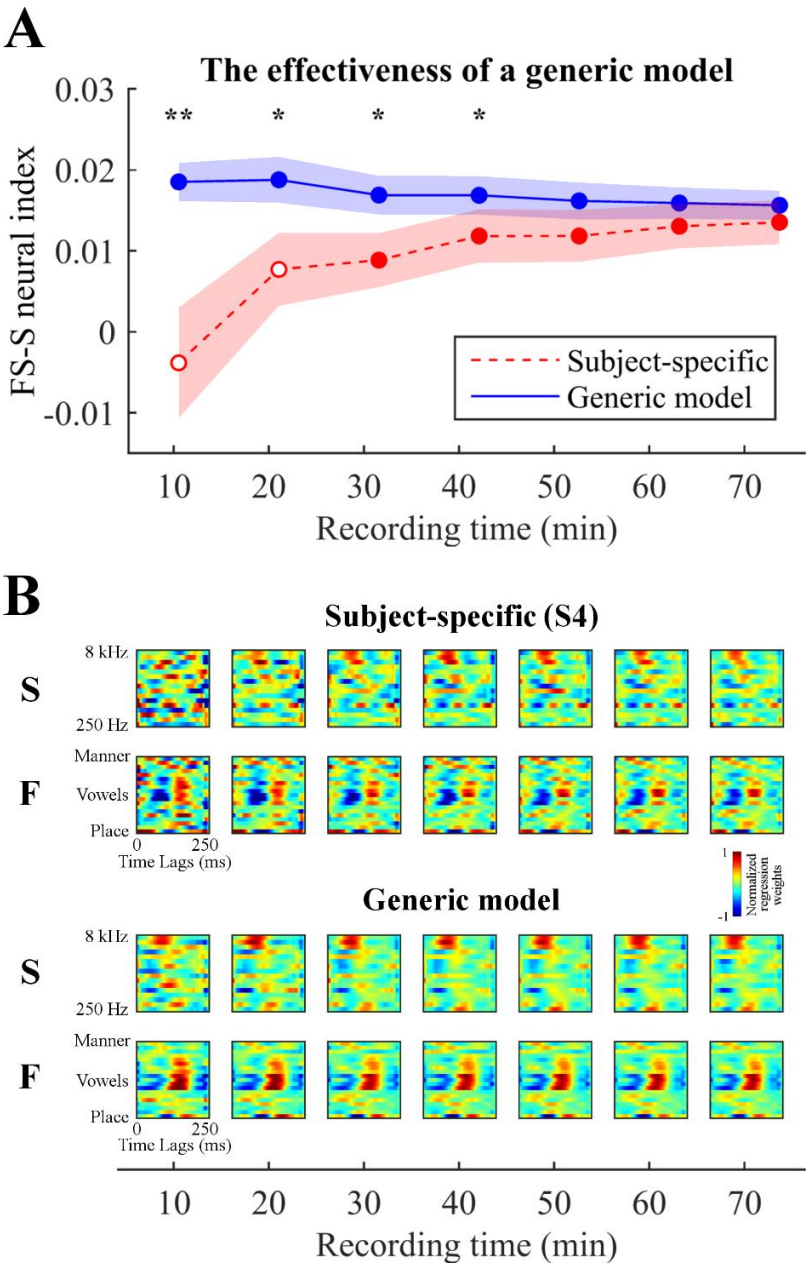


Figure 6.2: Generic models are effective for short recording times.
(A) The speech-specific neural index FS-S for increasing experimental time (minutes) is reported for the subject-specific models and compared with the result obtained using a generic model, i.e., trained on all subjects but the one used to evaluate the EEG prediction correlation. The subject-specific modelling approach needs at least 30 minutes of data to be sensitive to the FS-S effect, while the generic model produces significant results also for short experimental times. Importantly, a significant difference emerges between the two approaches for recording times under or equal to 40 minutes (** $p < 0.01$, * $p < 0.05$), which means that it's advantageous to use a generic model when little training data is available. Also, the use of speech-specific models does not improve the performance of a generic model, even when the whole 72 minutes dataset is used ($p > 0.05$). **(B)** The mTRF regression weights are shown for the S (Frequency vs time-lags) and F (Phonetic-features vs time-lags) models for increasing experimental time (minutes). Given a selected subject (S4, see Figure 6.1), this panel compares the mTRFs for its subject-specific model (fit on S4) and its generic model (fit on all the others). Smoothing of the x-axis was performed for visual purpose.

These results highlight the applicability of the generic modelling approach in the study of single subjects with limited amounts of data.

6.3.3 Sensitivity of EEG to phonetic features for limited recording time

As shown in Chapter 3, another possibly useful way of assessing speech encoding is to quantify how well the mTRFs to different phonetic features can be discriminated. But how much experimental data is needed in order to carry out these types of discriminative analysis? This is important because it determines which phonetic features have an impact on the model performances for different amounts of experimental data. Also, it determines which phonetic features can be further studied, for example to infer possible differences between subject groups. Here, we addressed this question by quantifying the discriminability between phonetic feature groups in the mTRF models. Specifically, unsupervised MDS was applied to the subject-specific phonetic-feature mTRF models using the 12 bilateral electrodes of interest. This approach allowed us to build a geometric space in which the Euclidean distance between phonetic features corresponds to the similarity of their neural responses. In this space, *k*-means clustering ($k = 2$) was performed to study the pairwise discriminability between feature groups (vowels, semi-vowels, fricatives, plosives), which was quantified by calculating the *F*-scores (the harmonic mean of precision and recall) between the actual grouping and the result of the clustering.

Figure 6.3 shows the evolution of the discriminability between phonetic features categories with the amount of experimental data. Chance level for the *F*-score measure (which changes for each pair of feature groups) was calculated by repeating this same procedure (MDS and *F*-Score) after randomly relabelling each phoneme occurrence and converting that into phonetic features. Each discriminability reported in **Figure 6.3** was obtained by subtracting the *F*-score derived when using the correct stimulus with its chance level (shuffled over 50 randomly relabelled versions of the stimulus). Individual subject values and their mean are reported in the figure.

In line with Chapter 3, EEG activity in response to vowels *(vow)* could be significantly discriminated from that to fricative (*fri*) and plosive (*plo*) consonants (paired Wilcoxon signed rank test, $p < 0.05$), while no significant difference between vowels and semi-vowels (*semi*) emerged. Importantly, these considerations were true even when only small amounts of experiment data (10 minutes) were available. The individual subject

data clarifies that, when enough data is available, the significant effects for the comparisons *vow-fri* (from 30 minutes of data) and *vow-plo* (from 50 minutes of data) correspond to an above-chance discriminability for every single subject. Also, EEG activity to vowels was more discriminable from plosive than it was from fricative consonants (this difference was significant for all experiment duration with the exception of 20 min; paired Wilcoxon signed rank test, $p < 0.05$). Interestingly, the discriminability of classes within consonants reveals a different pattern. In particular, semi-vowels required at least 20 minutes of experimental time to emerge as different from plosive consonants. Also, weak significant discriminability emerged between plosive and fricative ($p \lesssim 0.05$; with the exception of one data-point). In this case, a recording time of at least 60 (*plo-semi*) or 70 minutes (*plo-fri* and *fri-semi*) was required to achieve an above-chance result for every single subject.

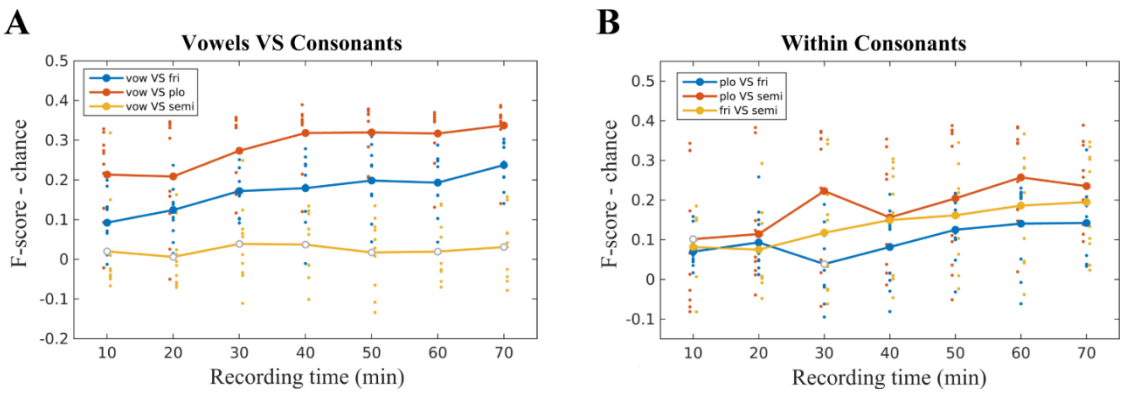


Figure 6.3: Sensitivity of EEG to phonetic features for limited recording time.
A measure of discriminability between phonetic features was derived from a MDS on the phonetic features mTRF model. In both figures, the x-axis indicates the amount of recording data and the y-axis reports a discriminability score (*F*-score). The comparison of each pair of feature-sets produced a distinct chance level, which was subtracted from the corresponding discriminability scores for visualisation clarity. Empty grey circles indicate non-significant discrimination values ($p > 0.05$). The small dots indicate the result on an individual subject level. **(A)** Vowels resulted discriminable from fricative and plosive consonants, and this difference emerged with 10 minutes of data. Vowels and semi-vowels were not significantly discriminable at any training time. **(B)** Plosive consonants and semi-vowels were significantly discriminated when at least 20 minutes of data was used. Similarly, plosive and fricative consonants are significantly discriminable for all recording durations, with the exception of 30 minutes.

6.4 Discussion

Language impairments are disorders that affect the understanding and/or use of spoken or written language, which carry the risk of poor social functioning, reduced independence and restricted employment opportunities (Clegg and Henderson, 1999; Paul, 2007; Reed, 2012). The disorder may involve the form of language (phonology, syntax, and morphology), its meaning (semantics), or its use (pragmatics), and includes

deficits such as specific-language impairment (SLI), aphasia, and dyslexia, among others. Early identification is crucial for improving long-term outcomes in many of these conditions, especially for early school age children who are less likely to have subsequent reading and academic problems in case of early diagnosis (Catts et al., 2002; Clark, 2010). In this context, the ability to derive noninvasively robust markers of natural speech processing at specific levels of the cortical hierarchy could be of great benefit for research in certain cohorts. Here we have investigated a number of practical considerations surrounding the novel framework for indexing the encoding of natural speech at the level of phonemes introduced in Chapter 3.

6.4.1 A methodological advance toward clinical application

This study demonstrates that a generic model is capable of indexing the cortical entrainment to several speech representations of interest. As one would expect, it was found that overall the EEG prediction correlations were lower for generic models than in the subject-specific approach (**Figure 6.1**). This is likely to be an effect of anatomical differences among individuals, which causes differences in the EEG signals between subjects. This subject-specific information would be lost when averaging between subjects, hence producing lower EEG prediction correlations. Even though the prediction values were smaller overall than in the subject-specific case, the generic modelling approach produces a similar pattern of prediction accuracies. Moreover, the generic model of a larger and reasonably homogeneous group would still encode cortical responses that are consistent across subjects. Potentially, this framework could benefit from such a larger dataset, and may require even shorter recording times to produce meaningful results. Here, out of the four speech representations used, the combination of acoustic and phonetic features (FS) is the best at predicting the EEG signal, while the envelope of speech is the worst. This result is relevant as the broadband envelope of speech has been used in several recent studies on auditory perception (Aiken and Picton, 2008; Nourski et al., 2009; Zion Golumbic et al., 2013a; Ding et al., 2014; Millman et al., 2015). And the generic modelling approach has been shown here to be able to produce a significant neural index of phoneme-level processing (FS-S).

The results discussed so far suggest that a generic modelling approach can be used to index cortical entrainment to phonetic features of speech. In order for this approach to be feasible for applied research in particular cohorts, this study aimed at assessing how

much experimental time it requires. **Figure 6.2** showed that subject-specific models are sensitive to recording duration and need at least 30 minutes of recording data to provide a significant index of phoneme-level activity (although, the more data the better), which limits the applicability of this framework. In this context, a solution is met by using a generic modelling approach, which was effective even with only 10 minutes of recording data. Furthermore, **Figure 6.3** demonstrates that phonetic-feature groups such as vowels, fricative consonants, and plosive consonants are discriminable already after 10 minutes of recording time (e.g., vowels VS fricative consonants, vowels VS plosive consonants). Unsurprisingly, the ability to separate phonetic-features increases with recording time, which highlights the importance of collecting as much experimental data as possible.

With the goal of minimising the experimental duration and facilitating clinical (and non-clinical) application, there are other considerations that are important to clarify. Firstly, the mapping procedure at the core of this framework (mTRF) is performed independently for each single electrode. In this context, Chapter 3 indicated a lack of topographical differences and that the strongest EEG predictability measures emerged from fronto-central scalp sites. The choice of focusing on a set of electrodes of interest in those sites also allowed us to investigate weaker effects that emerge at a group level, such as the sensitivity of EEG to specific phonetic features. However, no qualitative differences emerged between electrodes of interest and, in this sense, the effectiveness of this approach would not suffer from the reduction of the electrode-set, as long as the scalp areas of interest are used. This confirms that it is absolutely possible to obtain similar results by using only a few bilateral electrodes. However, the use of 16 or 32 scalp electrodes over the whole scalp surface may be important at the preprocessing stage, as it would facilitate artifact detection and channel interpolation to deal with noise and motor artifacts, which may be more problematic in specific cohorts (e.g., infants, older persons). Additionally, the use of a larger number of participants in each subject group may result in a further reduction of the amount of recording data needed to produce significant objective measures of speech perception.

6.4.2 Challenges and guidelines for real-world application

One possible shortcoming of the above generic modelling approach is that it relies on testing an individual subject using a model fit to other subjects. For studies comparing groups (e.g., typically developing children vs children with dyslexia), this means

combining data within each group to form separate generic models. This implicitly assumes a certain amount of homogeneity within each group, an assumption that is certainly problematic (Happe et al., 2006; Willems et al., 2016). In fact, **Figure 6.1** demonstrated that such variabilities affect the results even within the subject-group of this study. Intuitively, subjects with dynamics more similar to the group (subject-specific model similar to the correspondent generic model) will be characterised by higher EEG prediction correlations, while the opposite will happen for subjects with peculiar mTRFs. In this sense, the generic modelling approach could be used as a tool to investigate the homogeneity within a subject-group. This method is suitable to study within-subject effects, and indices of such effects (i.e., FS-S), properly normalised, could be used to compare different subject groups. In this context, interpretation of the analysis outputs needs to take into account the choice of subject-groups, as excessive within-group variability may hamper the fit of a generic model.

An alternative solution could be to define a unique generic model using a training control group, and to use such model to assess whether a new subject (e.g., a patient) belongs or not to the group. This approach has the advantage of having no limitations on the recording times for the training control group. However, the effectiveness of such an approach, which assumes some degree of homogeneity in each of the subject groupings, could not be verified here as it requires a dataset that includes at least two distinct subject groups. To this end, the measures of phonetic features discriminability (**Figure 6.3**) may provide a quantitative way to assess the homogeneity within a subject group. In particular, such measures were effective for every single subject when enough recording time was available and, for selected feature groups, 20 or 30 minutes of data were sufficient. However, it remains unclear how the particularities of the subjects used to fit a model will affect the predictions it produces for the test subject. Future work incorporating neuropsychological metrics and behavioural assays of speech perception will aim to clarify how this factor impacts our proposed methodology and to investigate the effectiveness of this framework at detecting specific processing problems at an individual subject level.

6.5 Summary

In summary, we have defined a framework to investigate speech processing using "direct" measures of cortical activity recorded with EEG. Importantly, the feasibility of applying this framework using shorter testing times was demonstrated. The approach provides a number of novel dependent measures of speech processing which can be used to assess speech processing on individual subjects in certain cohorts. In addition to an overall index of phonetic level processing (FS-S), we introduced a methodology to assess speech processing at the level of specific phonetic features, which may be important for investigating the causes and effects of specific speech and language disorders. For instance, dyslexia, which has been linked to phonological deficits (Goswami, 2015), may be related to altered/impaired processing of specific phonetic features. Another example is the study of language development. The processing of speech into phonetic categories is known to gradually develop through infancy and childhood (Kuhl, 2004); however, this has typically been investigated in the context of simple, discrete stimuli. The framework introduced here provides a new way to investigate such developmental processes in more naturalistic conditions.

Chapter 7. Deficits in right hemisphere encoding of natural speech correlate with psychometric measures of dyslexia

7.1 Introduction

The studies presented in Chapters 3 and 6 were oriented towards the development of a procedure to investigate the cortical processing of natural speech in particular cohorts, both clinical and nonclinical. The resulting procedure extracts objective electrophysiological measures of cortical tracking of specific acoustic and phonetic features of natural speech at both group and individual subject levels. This approach allows the extraction of reliable neural indices of speech using only 10 minutes of EEG recording, during which participants simply listen to natural speech from an audio-story. The present study explored the effectiveness of this approach to investigate speech processing in clinical populations while targeting one cohort in particular: Developmental dyslexia (hereafter dyslexia).

Dyslexia is a learning disorder defined by difficulties in the acquisition of reading despite an adequate learning environment and otherwise normal intellectual and sensory functioning (Snowling, 2000). The 5-10% of school aged children affected by dyslexia usually perform poorly in tests of phonological awareness, verbal short-term memory, and lexical-access, highlighting that this disorder affects the linguistic component involved in the process of learning to read (Vellutino et al., 2004). Furthermore, children with dyslexia have substantially higher rates of depression and anxiety, juvenile delinquency, school dropout, and lower chance of future employment (Sabornie, 1994; Wiener and Schneider, 2002; McNulty, 2003; Daniel et al., 2006; Baker and Ireland,

2007; Brooks, 2014). The reduction or eradication of these consequences could be achieved through a better understanding of this disorder. Specifically, the identification of the root causes of dyslexia could better inform the necessary conditions for environmental enrichment and allow for the development of clinical tools for its early diagnosis. This, in turn could benefit the education and future employment of millions of children (Goswami, 2015). However, a person with dyslexia accumulates far less experience in reading-related skills, such as language (e.g. vocabulary, phonological recording) and sensory processing (e.g. oculomotor control, visuospatial attention). In this context, it has been difficult to disambiguate causes from consequences of this developmental disorder, which are currently hotly debated.

Current theories propose that auditory, visual, or a modality-independent primary deficit underlies dyslexia (Goswami, 2015). The root cause of this disorder has been suggested to be phonological (Goswami, 2011; Richlan, 2012; Goswami and Leong, 2013; Lehongre et al., 2013; Clark et al., 2014; Goswami, 2015) and related to the temporal sampling of the auditory input (temporal sampling framework; Goswami, 2011). Another factor that may underpin such deficit and language impairments in general is working memory, which can be thought of as a temporary storage system necessary for a wide range of complex cognitive activities, including language processing (Baddeley, 2003). In addition, several studies with both adults and children found differences in amplitude and hemispheric bias between individuals with dyslexia and control groups (Lehongre et al., 2011; Giraud and Poeppel, 2012; Hamalainen et al., 2012; Poelmans et al., 2012). However, it has been argued that most of these studies did not hold the conditions necessary to assess causality (Goswami, 2015). Specifically, the account for control groups that match both age and reading-level was identified as crucial to disentangle causal effects from the consequences of reduced reading experience.

The ability to isolate cortical measures of speech processing at the level of phonemes could provide crucial insights to arbitrate the debate on the primary causes of this deficit which, in fact, have been suggested to be phonological. For this reason, dyslexia is an appropriate clinical application of the novel analysis approach introduced in Chapter 3 and further developed in Chapter 6. Here, we investigate causal effects of dyslexia by considering both age and reading-level matched control groups using measures of cortical entrainment to speech acoustics and phonetics derived with our approach. Children aged between 6 and 12 listened to an audio-story for 9 minutes while

non-invasive EEG data were recorded and, in a separate session, performed a battery of tests on language, memory, intelligence, and attention skills.

Under the hypothesis that dyslexia is underpinned by a phonological deficit (Goswami, 2011; Richlan, 2012; Goswami and Leong, 2013; Lehongre et al., 2013; Clark et al., 2014; Goswami, 2015), we expect such an impairment to affect measures of cortical entrainment that relate with phoneme-level processing, i.e. F, FS, and FS-S. Firstly, the present study aims to identify scalp areas that are affected by dyslexia. Secondly, the behavioural causes in such regions are investigated, with the goal of clarifying whether, where on the scalp, and at what latencies dyslexia is linked to a phonological deficit. The neural index FS-S allows us to isolate phoneme level processing (Chapters 3 and 6) and it has been shown to relate with intelligibility features of speech (Chapter 4). We hypothesised that FS-S would correlate with psychometric indices of phonological skills in scalp regions affected by dyslexia. The findings described in this chapter were presented at the Advances and Perspectives in Auditory Neuroscience (APAN) meeting as a poster (November 2016) and is currently in preparation as a research article as: "Deficits in right hemisphere encoding of natural speech correlate with psychometric measures of dyslexia", *in preparation*.

7.2 Material and methods

Sixty-seven children (twenty-six female) aged between 6 and 12 years old, whose native language is English, participated in the experiment. The ethics committee for Human Research at Western Sydney University approved all the experimental methods used in the study. Informed consent was obtained from the parents of all the participants. Children also gave verbal assent for the study.

7.2.1 Subjects

Children were grouped into: 25 participants with dyslexia (DX; 8 female) and 42 control subjects (CTR; 18 female). Children were recruited in Sydney via advertisements in local media or via a database of families who previously expressed interest to participate in infancy and child research. All participants reported having no hearing difficulties. Families' socio-economic statuses were calculated based on the average household weekly income of their area of residence (Australian Bureau of Statistics). All families came from middle or higher middle socio-economic backgrounds.

An additional grouping was derived to compare a DX group with two control groups: one matched by age (CA) and the other by reading-level (RL). DX, CA, and RL were composed of 13, 13, and 32 children respectively. The additional 14 children were not included as their reading-levels were either too high or too low compared to the rest of the subjects.

7.2.2 Behavioural measurements

Group assignment (DX vs CTR) was determined based on the children's performances on tests from a screening battery that included measures of language, memory, intelligence, and attention. Children were assigned to the DX group if A) they obtained a score of 1.5*SD (standard deviation) below the age-appropriate mean in one or more phonological processing task and at least one reading task, and if B) had average scores (within 1*SD from the age appropriate mean) on the grammatical competence tests, and C) had average non-verbal IQ score and no indications of Autism Spectrum Disorder (ASD) or attention deficit hyperactivity disorder (ADHD). Children were assigned to the control group if they obtained average scores (within 1*SD) on all the tasks of the screening battery and had no indications of ASD or ADHD. The specific psychometric measures are described below:

Word and non-word reading: The sight word efficiency and the phonemic encoding efficiency sub-tests of the Test of Word Reading Efficiency (TOWRE; Torgesen et al., 2012) were administered. The TOWRE consists of two lists, one of 66 words and another of 66 non-words. In two separate trials, children are required to read as many items as possible from each list in 45 seconds. A standardised score (M = 100, SD = 10) is computed based on how many words are read accurately in this time for each test.

Phonological processing: (1) Phonological awareness: Four sub-tests of the phonological awareness battery of the Comprehensive Test of Phonological Processing (CTOPP; Wagner et al., 2013) were administered. A) Elision – children were required to pronounce a word while omitting one of its component sounds, e.g. "say cup without /k/". B) Blending words – children were required to hear two parts of a word and were asked to combine them and produce the resulting word, e.g., "/pen/ and /səl/ make pencil". C) Sound matching – in this test, children saw two images of objects and were required to

point to the object whose label contained a target sound, e.g., when shown the objects sun and ball, the child is asked to show the one that starts with /s/. D) Phoneme isolation – children are required to listen to a word and identify one of its component sounds, e.g., "what is the second sound of the word train". A composite standardised score for phonological awareness is then computed (M = 100, SD = 10). (2) Phonological short-term memory (PTSM): all children completed the digit and non-word repetition subtests of the CTOPP (Wagner et al., 2013). Children were presented with sequences of digits or non-words that increased with complexity after each trial and were required to repeat them in the same order as they were presented. This yields a composite standardised score for phonological memory (M = 100, SD = 10).

Rapid Symbolic Naming: The rapid digit naming and rapid letter naming subtests of the CTOPP were administered. Children were presented with a list of 36 items (digits or letters respectively) on a card and required to name as many as possible in the period of 2 minutes. The number of accurately named items in that time is used to calculate a standardised composite rapid symbolic naming score (M = 100, SD = 10).

Working memory: Children completed the forward and backward number repetition subtests of the Clinical Evaluation of Language Fundamentals test (CELF; Semel et al., 2006). A composite standardised working memory score was obtained based on the number of items that the child could successfully recall in each subtest (M = 10, SD = 3).

Expressive vocabulary: The expressive vocabulary subtest of the Wechsler Individual Achievement Test (WIAT; Wechsler, 2009) was completed. The experimenter showed children an image and described it (e.g., "tell me the word that means a brush for cleaning teeth"), and children were required to name the image. The test is discontinued after 4 consecutive incorrect answers or when the entire set of 17 items is completed, yielding a standardised expressive vocabulary score (M = 100, SD = 10).

Grammatical competence: The Test of Reception of Grammar (TROG; Bishop, 2003a) and the Recalling Sentences subtest of the CELF (Semel et al., 2006) were administered. In the TROG, children were shown a card with four images and heard a sentence. They were required to point to the image on the card that was described by the sentence. The total number of correct responses is used to calculate the standardised

reception of grammar score (M = 100, SD = 10). In the Recalling Sentences subtest, children heard a sentence and were required to repeat it verbatim. Responses are scored according to the number of errors made in each repetition, and used to compute a standardised score for this subtest (M = 10, SD = 3).

Non-Verbal Intelligence: Children complete the matrices subtest of the Kaufman Brief Intelligence Test (KBIT; Kaufman and Kaufman, 2004). The number of matrices completed correctly out of a maximum of 46 items is used to compute a standardised non-verbal intelligence score (M = 100, SD = 10).

Parental questionnaires: In addition to the screening battery, children's parents completed the Children's Communication Checklist (CCC-2; Bishop, 2003b) and the Swanson, Nolan, and Pelham rating scale (SNAP-IV; Swanson, 1992). The CCC-2 is used to assess children's general communicative abilities and identify communicative deficits characteristic of SLI or Autism Spectrum Disorder (ASD). The SNAP-IV is used to identify behavioural patterns characteristic of ADHD or other behavioural disorders. No children who were included in the final sample showed any indications of ASD or ADHD.

7.2.3 EEG Experimental Procedure

Participants were presented for 9 minutes with an audio-story read by a female Australian English speaker. The stimulus was presented monophonically at a sampling rate of 44,100 Hz using loudspeakers while subjects watched the corresponding cartoon. Note that the visual input only generally matched the events narrated in the audio-story, whereby they found correspondence in the general meaning but not in the detailed temporal events such as speech. High density EEG was recorded using 129-channel Hydrocel Geodesic Sensor Net (HCGSN), NetAmps 300 amplifier and NetStation 4.5.7 software (EGI Inc.) at a sampling rate of 1 kHz with the reference electrode placed at Cz. The electrode impedances were kept below 50 kΩ.

7.2.4 EEG Data Preprocessing

Data were analysed offline using MATLAB software (The Mathworks Inc.). EEG electrodes that were positioned at the jaw, mastoids, and forehead were removed from the

analysis because of their excessive noise. EEG signals from the remaining 93 channels were digitally filtered in three frequency-bands: delta-band (1-4 Hz), theta-band (4-8 Hz), and 1-8 Hz, using Chebyshev Type 2 filter in both the forwards and backwards directions to remove phase-distortion. Data were referenced to the average of all channels. In order to reduce processing time, data were down-sampled to 100 Hz. EEG channels with variance that exceeded three times that of the surrounding channels were labelled as bad channels, and replaced by an estimate calculated using spherical spline interpolation (EEGLAB; Delorme and Makeig, 2004).

7.2.5 Model Evaluation

As performed in Chapter 3, a model-based analysis was conducted to quantify how well the EEG signal reflects the encoding of different features of speech (**Figure 7.1A**). Specifically, the speech representations used were:

- Spectrogram **(S)**;
- Phonetic features **(F)**;
- A combination of F and S **(FS)**.

Please refer to Section 3.2.2 for a detailed description of these speech representations.

The idea is to fit linear mTRF models to describe the forward mapping from a speech representation to the corresponding EEG and then to test that model by seeing how accurately it can predict EEG from a new trial. This is quantified using Pearson's correlation and a 9-fold cross validation. Because only 9 minutes of data were available for each subject, prediction correlations were derived using a generic modelling approach, as defined in Chapter 6. Such an approach entails the averaging of mTRFs within specific subject groups and, therefore, it assumes a certain degree of homogeneity for each group.

The analysis was conducted using two different approaches that use generic models:

a) *Within-group approach* (**Figure 7.1B**): Given two subject groups A and B, mTRF models are trained for every subject in each group and EEG prediction correlations are calculated using a cross-validated generic modelling approach within each group. A and B can be compared both in terms of their mTRFs and prediction correlation values. This approach is used to compare DX with CA and RL at each EEG electrode.

b) *Cross-group approach* (**Figure 7.1C**): This approach originates from an idea that was mentioned in Section 6.4. The main rationale is to derive a template

using a sample from a group of subjects, and then to test whether a new subject belongs or not to that same group. The control subjects group (CTR) is randomly split into two partitions of equal size: CTR-1 and CTR-2. mTRF models are fit for each subject in CTR-1. A single generic model is obtained by averaging models for all subjects in this group. This model is then used to predict the EEG of all other subjects (CTR-2 and DX). The rationale is that CTR-2 should be homogeneous with CTR-1, therefore the EEG responses of its subjects should be predictable. Differently, a group like DX may be underpinned by a different response pattern, which would make the EEG less (or not) predictable using a model trained on control subjects. The procedure is then repeated using CTR-2, providing prediction values for CTR-1. Prediction values for non-control subjects using models fit on CTR-1 and CTR-2 were averaged. Mean prediction accuracies were derived by averaging all EEG electrodes. Results were averaged over 100 repetitions of this procedure, which used random binary partitioning of the CTR group. Correlations between these EEG predictability indices and psychometric behavioural measurements was evaluated using the Pearson's correlation index. This analysis was also conducted at each individual scalp electrode.

In both approaches, the model time-lags and the regularisation parameter λ were optimised as in Chapter 3.

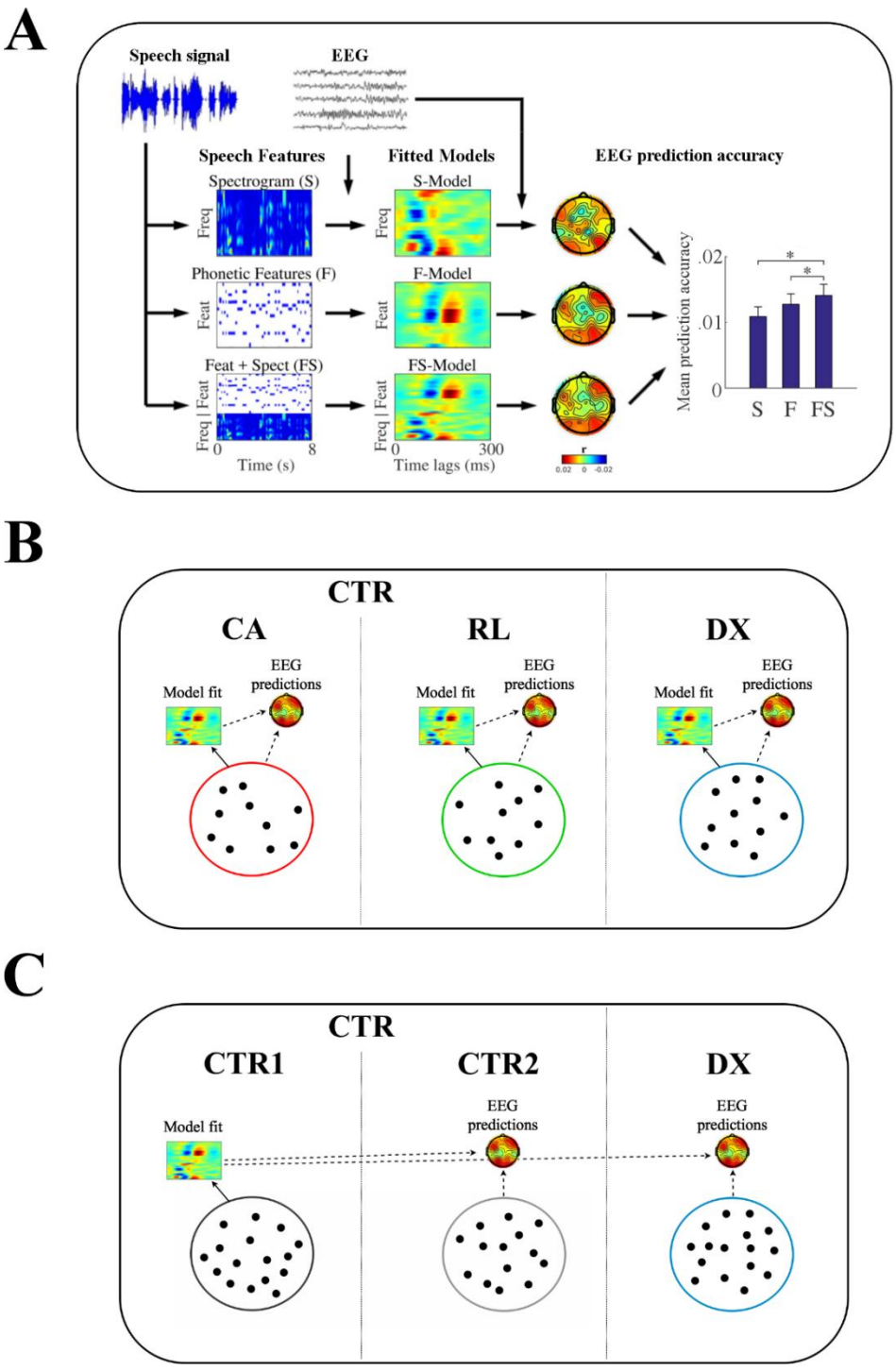


Figure 7.1: Assessing the encoding of speech features in EEG.
(A) 128 channel EEG data were recorded while subjects listened to continuous, natural speech consisting of a female speaker reading a story. We used linear regression to fit multivariate temporal response functions (mTRFs) between the low-frequency (1-8 Hz) EEG and five different representations of the speech. Each mTRF model was then tested at each scalp location for its ability to predict EEG using leave-one-out cross-validation and Pearson's correlation measures. **(B)** *Within-group approach*: EEG prediction accuracies for a subject are calculated using mTRF models fit using the other subjects of the same group. This approach allows to compare models and EEG predictions between groups. To this end, comparisons with the dyslexia group (DX) were performed for control groups (CTR) that were matched either by age (CA) or reading-level (RL). **(C)** *Cross-group approach*: Control subjects were randomly partitioned into 2 groups (CTR1 and CTR2). EEG prediction accuracies for a subject (either in CTR2 or DX) are calculated using the mTRF models fit using a sample set of control subjects (CTR1).This procedure was then repeated for a model fit on CTR2.

7.2.6 Statistical Analyses

Unless otherwise stated, all statistical analyses were performed using non-parametric Wilcoxon signed rank tests, while non-paired comparisons were conducted using two-sided Wilcoxon rank sum tests. For tests involving all scalp electrodes, a cluster-based non-parametric analysis was conducted, with two as the minimum cluster size (Maris and Oostenveld, 2007). This statistical test takes into consideration the scalp distribution of the measure of interest by performing a permutation test on the cluster of electrodes with highest score i.e., the most important cluster according to the metric of interest. This approach provides a solution to the multiple comparison problem by including biophysical motivated constraints that allow to increase the sensitivity of this statistical test in comparison with a standard Bonferroni correction. Also, the false discovery rate (FDR) method (Benjamini-Hockberg-Yekutieli procedure; Benjamini and Yekutieli, 2001) was used to assess significance in analyses that involved multiple comparisons of a neural index with the several psychometric measures.

7.3 Results

7.3.1 Reduced EEG predictability in dyslexia

A *within-group* mTRF analysis was conducted on DX, CA, and RL. This approach compares subject groups in terms of how well acoustic and phonetic features of speech are reflected by the EEG signal. Measures of EEG predictability (Pearson's *r*), which quantify the cortical entrainment to the corresponding features of speech, were derived for each speech representation. Qualitatively similar patterns emerged for the two control groups, while noticeably different scalp distributions resulted for subjects with dyslexia (**Figure 7.2A**). The neural indices, averaged across electrodes, showed an overall reduction for dyslexia compared to both control groups for all speech models (Wilcoxon Rank-Sum Tests: DX versus RL, $p = 0.027$, $p = 0.014$, $p = 0.009$ for S, F, and FS respectively; DX versus CA, $p = 0.027$, $p = 0.011$, $p = 0.021$ for S, F, and FS respectively). **Figure 7.2B** reports scalp areas that showed significant differences ($p < 0.05$) between DX and the control groups.

Effects in the EEG delta-band involved broader scalp areas than in theta-band ($\text{significant-area}_{\text{delta-band}} / \text{significant-area}_{\text{theta-band}} = 2.64$, 1.49, and 11.33 for S, F, and FS respectively, where area was calculated as {DX versus CA} ∪ {DX versus RL}),

nevertheless the combination of both bands (1-8 Hz) produced the strongest results. Specifically, differences between DX and both control groups, which reflect effects of dyslexia that are not due to reduced reading skills (Goswami, 2015), were most pronounced for the FS-model in the combined 1-8 Hz band. Such effects emerged in three scalp regions (corresponding to the areas coloured in yellow in **Figure 7.2B**): a frontal area ROI1; a region in the right hemisphere ROI2; and an occipital area ROI3. Significance also emerged for an isolated electrode in the posterior-right region, which was not associated to any ROI as a minimum cluster size of 2 was used for the cluster statistics. ROI2 can be further divided into two clusters, one more frontal (ROI2a) and the other central (ROI2b), which exhibit significant suppression and enhancement due to dyslexia respectively. Furthermore, significant effects of dyslexia using the FS and F speech representations involved broader scalp areas than for the purely acoustic representation S (area(FS) / area(S) = 3.52; area(F) / area(S) = 2.61).

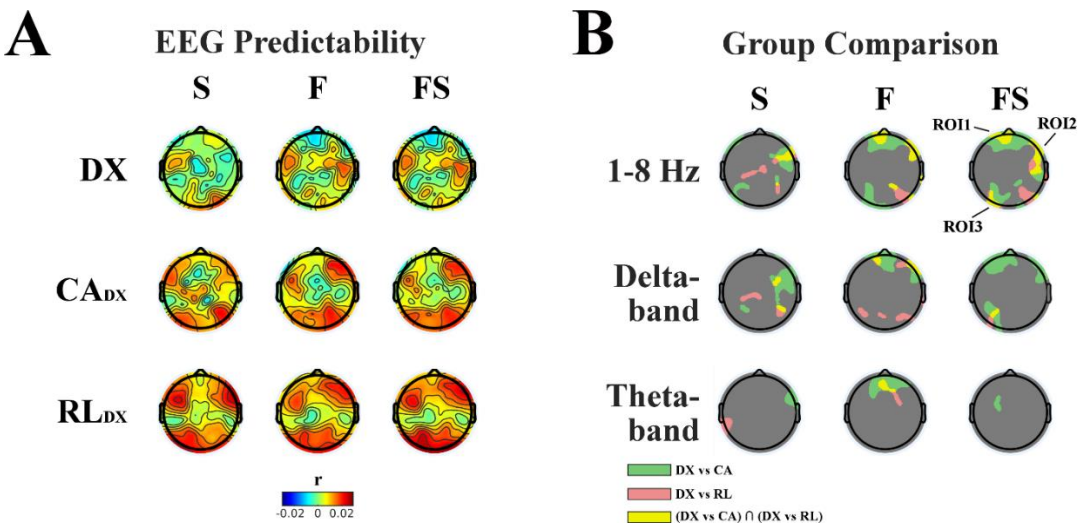


Figure 7.2: Right biased reduction of cortical entrainment to natural speech in dyslexia.
(A) EEG (1-8 Hz) prediction correlations (Pearson's *r*) averaged across participants are reported here for each model (acoustic – S, phonetic – F, combined – FS) and for each subject group (dyslexia – DX, age matched control – CA, reading-level matched control – RL). **(B)** EEG predictability was compared between DX and the two control groups, for each model and scalp electrode using a paired Wilcoxon test. Coloured areas indicate EEG channels that showed significant effects between subject groups (cluster-based nonparametric statistics, $p < 0.05$). Direct effects of dyslexia were disentangled from other indirect effects (such as contrasts due to differences in reading skills or age) by identifying electrodes that showed significant differences between DX and both control groups ({DX – CA} ∩ {DX – RL}), which are indicated in yellow. The group differences emerged for the EEG frequency-band 1-8 Hz were shown to be mainly (but not uniquely) driven by effects in delta-band.

7.3.2 Cortical entrainment to speech correlates with linguistic skills

The previous section showed that cortical entrainment to acoustic and phonetic features of natural speech is affected by dyslexia, and that this is not due to differences in age nor reading skills. However, the precise factors that underpin these effects remain unclear. Specifically, are they related to a phonological impairment, or do they reflect some other factors, such as an attentional deficit? We sought to answer to this question by assessing correlations between EEG predictability scores and psychometric measures.

A *cross-group* approach was used to test this hypothesis. **Table 7.1** reports correlation values between EEG predictability using FS (after averaging across all scalp electrodes) and all behavioural scores. Significant positive correlations emerged with measures of 'phonological awareness, memory, and rapid symbolic naming', 'digit span', 'recalling sentences', 'reception of grammar', and 'reading non-words' (FDR corrected, $\alpha = 0.05$). Note that, differently from the *within-group* analysis in **Figure 7.2**, this analysis does not exclude potential effects of reading-skills (an issue that will be addressed in the following section). In fact, *cross-group* analyses were performed on all subjects rather than only on the subsets DX, CA, and RL.

The same correlation analysis was conducted on the composite EEG predictability measure FS-S, which was suggested in the previous chapters to represent an isolated index of phonological processing. In this case, this neural measure correlated with behavioural scores of 'phonological memory and rapid symbolic naming', 'digit span', 'reading words and non-words', 'general communication', and 'inattention'. These various patterns of results were only partially expected. In fact, the possible link of FS and FS-S with phonological processing was hypothesised to translate into strong correlations with measures of phonological skills. While this was certainly the case for the FS index, the significant correlations with reading skills and inattention hamper the interpretation of this result. Nevertheless, this analysis indicates that there is a link, either direct or indirect, between EEG predictability measures and standard psychometric measures of language skills.

Category	Psychometric Test	Correlation with FS	Correlation with FS-S
Phonological Skills	Phon. awareness	0.338	0.281
	Phon. Memory	0.462	0.340
	Rapid symbolic naming	0.348	0.334
Language Fundamentals	Digit span	0.324	0.358
	Recalling sentences	0.407	0.280
	Reception of grammar	0.344	0.207
Reading Efficiency	Words	0.251	0.358
	Non-words	0.359	0.405
General communication	Children's communication checklist	0.271	0.336
IQ	Brief intelligence test	0.296	0.230
ADHD	Inattention	-0.132	-0.359
	Hyperactivity	-0.085	-0.294

Table 7.1. Cortical entrainment measures correlate with measures of phonological and language skills.
EEG predictability measures for the FS-model and the ones derived as FS-S (Pearson's *r*) were averaged across all electrodes and then correlated with psychometric measures of phonological skills, language, intelligence, and attention. Significant values are reported with black font (FDR correction, $\alpha = 0.05$).

7.3.3 Topographic specificity of the effects of dyslexia

The correlations in **Table 1** were obtained by averaging the neural measures across all scalp electrodes. While this analysis confirmed that such neural indices are linked to language skills, it could not disambiguate effects of phonological skills from other measures of language, reading, and attention. In this context, it is possible that effects of dyslexia that arise at distinct scalp locations are underpinned by different factors, which would provide an opportunity to disentangle the effects due to phonological skills. In particular, we hypothesised that electrodes that were sensitive to the effects of dyslexia (**Figure 7.2B**, ROI1-3) would show correlation with psychometric measures reflecting the causes of such effects. Specifically, this analysis was aimed at revealing whether dyslexia is linked to a phonological deficit.

Figure 7.3 reports the correlation values of cortical entrainment measures for FS-S, FS, S, and F with six most relevant psychometric measures of phonological, reading, and memory skills. Firstly, the FS-S metric produced significant positive correlations with the measures 'phonological awareness' and 'phonological memory' that emerged

only in the right hemisphere and that covered entirely ROI2. ROI2 showed also a correlation between FS-S and 'Digit span', a measure that has been linked to both short- and long-term memory (Jones and Macken, 2015). Finally, no significant correlation with 'reading words', 'reading non-words', and 'sentence recall' emerged in any ROI.

The same correlation analysis was conducted for FS, F, and S. EEG prediction scores for the F-model, which is based on phonetic features, were similar to the results for FS-S, with the exception of 'phonological awareness' that did not show a significant correlation in ROI2. The S-model, which is based on speech acoustics, did not show significant correlations with psychometric measures in any ROI, with the exception of a single electrode in ROI2. In addition, as hypothesised, FS scores positively correlated with 'phonological awareness and memory' in ROI2, and also with 'sentence recall' and 'digit span'. These results may suggest a major link between phonological and memory skills with dyslexia, involving in particular the right hemisphere. Importantly, these effects are not due to differences in reading skills.

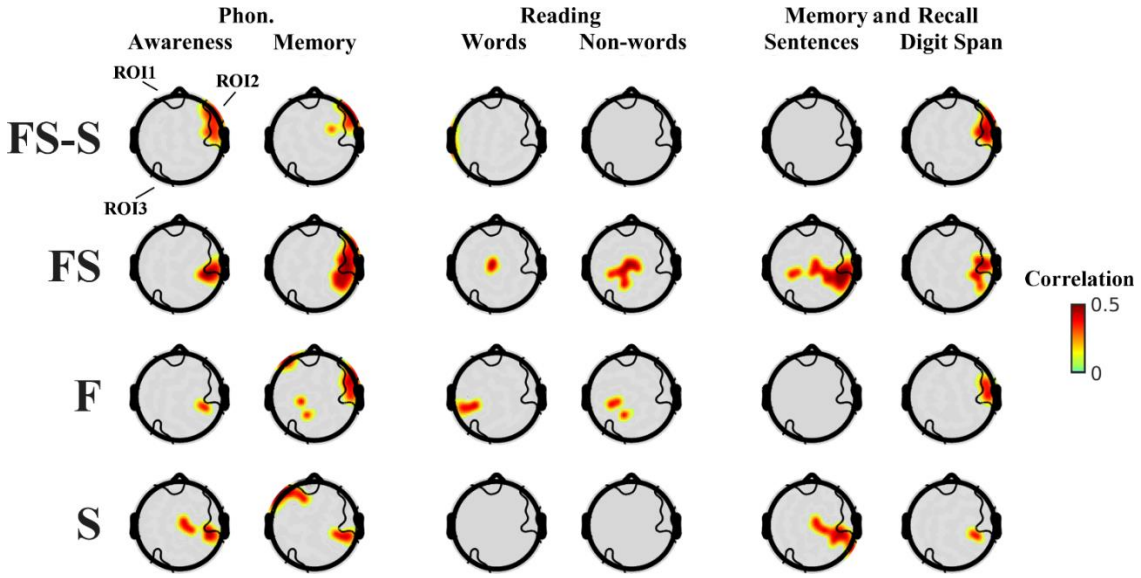


Figure 7.3: Topographic specificity to language skills of the effects of dyslexia.
Correlations (Pearson's *r*) between EEG predictability measures and all psychometric measures are reported for all scalp locations using all subjects. Significant correlations were identified by using cluster-based nonparametric statistics ($p < 0.05$; FDR correction for multiple comparisons was applied for each electrode and neural index, $\alpha = 0.05$). Grey areas indicate scalp areas that did not reach significance. Scalp areas that showed a significant effect of dyslexia for the FS-model (see Fig. 7.2B, 1-8 Hz) are indicated by black contours (ROI 1-3). The neural measure FS-S, which was suggested to represent an isolated index of phoneme-level processing, correlates with behavioural measures of phonological skills specifically for ROI2.

Indeed, one issue is that several of the psychometric measures used are highly correlated with each other (**Figure 7.4**). For this reason, despite the use of cluster statistics that includes correction for multiple comparisons, the correlation analysis is prune to

spurious correlations, which challenges the identification of the causes of dyslexia. For example, ROI2 showed correlations with both measures of 'phonological awareness' and 'digit span'. However, the two psychometric indices are correlated with each other (r = 0.41, p = 0.0003), thus it is difficult to conclude that ROI2 is linked to one, the other, or both skills. On the other hand, spurious correlations do not always occur as an effect of the pair-wise correlations between behavioural measures shown in **Figure 7.4**. For example, the 'IQ' score correlated with most measures of language skills, however it showed no significant interactions with any of the neural measures of interest (**Figure 7.5**). To further reinforce the present results, the individual electrode correlations between behavioural and neural measures in **Figures 7.3** and **7.5** were recalculated by using a partial correlation that excluded possible interactions with IQ and age of the participants. This less conservative analysis confirmed the correlation between FS-S with 'phonological awareness' and 'digit span' skills in ROI2, while correlations with FS in that same region emerged only for the behavioural measure 'recalling sentences'.

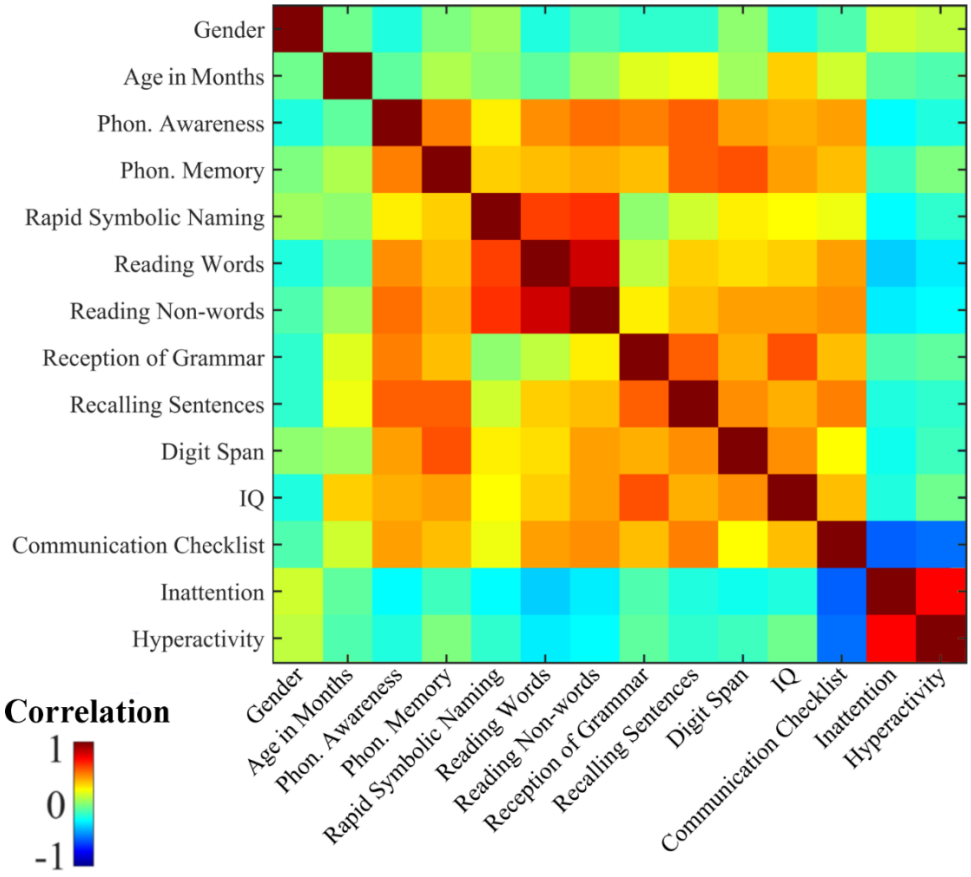


Figure 7.4: A further look at the psychometric measures.
Correlation (Pearson's *r*) between all pairs of psychometric measures are reported

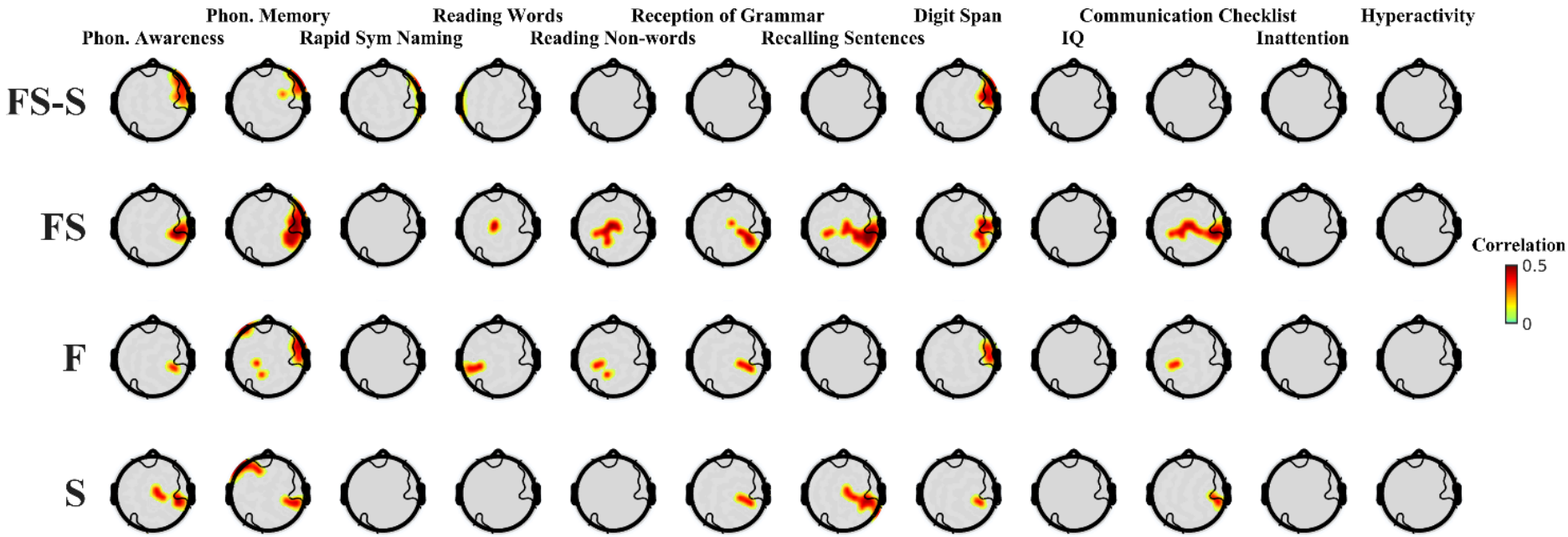


Figure 7.5: Topographic specificity to language skills of the effects of dyslexia.
Correlations between EEG predictability measures and all psychometric measures are reported for all scalp locations using all subjects. Significant correlations were identified by using cluster-based nonparametric statistics ($p < 0.05$; FDR correction for multiple comparisons was applied for each electrode and neural index). Grey areas indicate scalp areas that did not reach significance. Scalp areas that showed a significant effect of dyslexia for FS (Figure 7.2) are indicated by black contours.

7.4 Discussion

Chapters 3 and 6 introduced a novel methodology to study natural speech perception at the level of phonemes. This chapter represents the first clinical application of such methodology and targeted children with developmental dyslexia, a deficit whose root causes remain uncertain and have been suggested to lie in impaired phonological processing (Goswami, 2011; Richlan, 2012; Goswami and Leong, 2013; Lehongre et al., 2013; Clark et al., 2014; Goswami, 2015). Specifically, the deficit may relate to the cortical mechanism of temporal sampling and cortical tracking to speech (Thomson and Goswami, 2008; Goswami, 2011). To test this hypothesis, we conducted our novel EEG analysis on children aged between 6 and 12 years that were presented with 9 minutes of natural speech from an audio-story.

Firstly, a group level analysis demonstrated that dyslexia affects the cortical tracking of speech. Dyslexia produced an overall reduction in cortical tracking of speech across all scalp electrodes. Analysis at individual scalp electrodes revealed significant effects of dyslexia specifically in three distinct ROIs. As hypothesised, results at the individual subject level revealed positive correlations between these effects and measures of phonological skills and memory capacity, supporting the theory of impaired phase locking to speech in dyslexia. Furthermore, these results demonstrate the effectiveness of this experimental framework in studying speech perception in clinical (and other) cohorts of interest.

7.4.1 Impaired phase-locking to speech features in dyslexia

In line with the initial hypothesis, cortical entrainment to acoustic and phonetic features of speech were affected by dyslexia. As discussed throughout this thesis, EEG predictability constitutes a reliable indicator of cortical entrainment to the specific features of speech used at the model fit and prediction stages. In this context, the overall reduction in EEG predictability measures in DX may reflect impairment in the cortical mechanisms of phase-locking to speech. An alternative explanation for this reduction is a lower homogeneity in the EEG responses within the DX group compared to CA and RL. This would affect the analysis procedure at the model fit stage, which requires averaging across subjects. Nevertheless, the significant effects of dyslexia were not

uniform across the scalp and emerged as a consequence of both increased and decreased EEG predictability (**Figure 7.2**), suggesting that dyslexia affects the topographic patterns of phase locking rather than just its overall magnitude. For this reason, we contend that this result reflects a deficit in phase locking due to dyslexia that selectively affects specific cortical areas, rather than simply producing an overall reduction of cortical tracking.

7.4.2 Atypical phonological processing in right hemisphere in dyslexia

The correlation analysis in **Figures 7.3** and **7.5** demonstrated a variety of scalp distributions for different psychometric measures and indicated a main role of phonological and memory skills in the same ROIs that showed to be affected by dyslexia. Indeed, that EEG predictability correlates with psychometric measures is not surprising *per se*. In fact, both electrodes in ROI1-3 and the chosen psychometric measures (used for diagnosis) are affected by dyslexia by definition, which could determine correlations between these neural and behavioural indices. Instead, this analysis was conducted to identify which specific psychometric measures were generating such differences and whether these exhibited topographic specificity.

The hypothesis of a phonological deficit in dyslexia related to the temporal sampling mechanisms finds support in the strong correlations emerged here between the behaviourally measured deficit in phonological skills and neural indices of cortical entrainment to phoneme-level speech. In particular, this result emerged for FS-S in ROI2, indicating a right hemisphere specificity of this deficit. This result is in line with recent literature on speech perception that hypothesised hemispheric specialisation to speech features at different rates (Poeppel, 2003). According to this processing scheme, named "asymmetric sampling in time" (AST) hypothesis, a deficit in the processing of specific speech units, such as phonemes, would affect speech processing differently between the two hemispheres. This expectation was met by previous studies that showed effects of dyslexia in the right hemisphere (Goswami, 2011; Peter et al., 2016) and on hemispheric asymmetry (Heim et al., 2003; Abrams et al., 2009). Given the correspondence between the present results and previous findings, we contend that temporal sampling of phonological units is altered by dyslexia and that this phenomenon affects specifically the cortical entrainment to phonological units in the right hemisphere.

7.4.3 Neural correlates of reduced working memory and reading level in dyslexia

Together, these results confirm the hypothesis that motivated this work, but they do not present a simple story. Indeed effects of dyslexia, that were not due to differences in reading-level, emerged in right hemisphere (ROI2) and related specifically to a phonological deficit. However, EEG predictability in the same scalp area correlated with measures of working memory ('digit span' and 'recalling sentences') and, in minor part, with other linguistic skills (e.g. 'communication checklist'). Of course, it is possible that some of these correlations are spurious, thus indicating only indirect links between cortical oscillations and such behavioural scores. Furthermore, as expected, measures of attention and IQ did not show any interaction with the atypical cortical entrainment to speech features.

The behavioural measures 'communication checklist' and 'recalling sentences' showed some significant correlations only with the neural indices FS and S in ROI2. However, the effect involved only part of ROI2, suggesting that it is either spurious or simply a minor contributor to the effect of dyslexia in that scalp region. Differently, strong correlation values emerged for the measure of working memory 'digit span' with FS-S, FS, and F specifically in ROI2. As for measures of phonological skills, ROI2 was involved only for speech models that included phonetic features, suggesting once again the link of that region with phonological processing.

Interestingly, no correlations emerged for ROI1 and ROI3 from the individual electrode analysis. Indeed, this may be the consequence of a limited sensitivity of this analysis, possibly because of the low SNR of the recordings, or it may entail that those regions are affected by multiple complementary factors that do not emerge for pair-wise interactions. One attempted explanation is that the effects seen for the occipital region ROI3 relate to differences in the processing of visual cues. However, this is unlikely because auditory and visual inputs were related semantically but not at the level of speech acoustics and phonetics, while the TRF analysis regressed out specifically responses to such lower-level features of speech. Another explanation could relate to the fact that a group difference does not necessarily reflect a linear gradient corresponding to some behavioural measures of language skills that would produce strong correlation values. For example, a sharper sigmoid-like relationship could underpin such a relationship. However, post-hoc investigation using multiple kernels would be too sensitive to false-positives, thus future studies will require precise prior hypotheses on such interactions.

Other significant correlations emerged for electrodes not belonging to any ROI. This was the case for the reading level measure, indicating that this factor is not a main contributor to the atypical cortical entrainment in dyslexia, which reinforces previous studies on dyslexia and cortical entrainment that did not include a reading-level matched group. Nevertheless, this chapter confirmed the strong relationship between reading-level and the cortical processing of speech, hence we believe that future research should include such a control group when investigating dyslexia and its root causes.

7.5 Summary

Our results provide, for the first time, evidence for impaired low-frequency neural entrainment to phonetic features of natural speech in dyslexia. Furthermore, this phenomenon exhibited a right hemispheric bias, which was associated specifically with a deficit in phonological skills and was disentangled from differences in reading-level. This finding provides quantitative support for the temporal sampling framework formulated by Goswami (Trends Cogn Sci, 2011). Also, the neural indices of cortical entrainment correlated with psychometric measures of language skills, and FS-S correlated specifically with measures of phonological skills in scalp regions affected by dyslexia. This suggests that these neural indices may provide an objective metric to investigate language development and impairment.

Chapter 8. General Discussion

Speech is central to human life. However, the neural mechanisms that allow speech comprehension are poorly understood. As we know from Chapter 2, one major challenge lies in capturing the high spatial and temporal detail of what is thought to be a hierarchical system. The ability to study cortical dynamics at specific hierarchical levels in isolation may be key in:

A) Understanding the cortical underpinnings of speech comprehension in healthy adults and typically developing children;
B) Investigating the causes of deficits involving speech and language functions.

While new non-incremental insights on this system have been produced by ECoG and fMRI work, these have been constrained in terms of population sample and cortical coverage of study, and by low temporal resolution respectively.

This thesis introduces a novel framework to isolate neural activity at specific levels of the speech processing hierarchy using natural stimuli and high temporal resolution. This EEG-based approach, which can be extended to MEG, is applicable to a wide variety of subjects, from infants to older persons, and in both clinical and non-clinical populations.

The first part of this thesis, consisting of Chapters 1-3, motivates this research (Chapter 1) and provides a summary of the recent work in the field (Chapter 2). Chapter 3 defines the novel framework and demonstrates that it is sensitive to the cortical processing of speech at the level of phonemes. The latter is considered a crucial step of speech comprehension, as it represents an early level of categorical perception that is "special to speech", i.e. that occurs specifically for speech and not for other non-speech sounds (Mesgarani et al., 2014; Leonard et al., 2016).

The second part of the thesis, composed of Chapters 4 and 5, deals with the key point A described above. The EEG-study of Chapter 4 isolates quantitative indices of phoneme-level processing to investigate the mechanisms of integration of prior knowledge with sensory input on healthy adults during speech perception. Chapter 5 provides further insights on that same cortical process using an MEG source-space analysis with a focus on disentangling neural activity from different cortical sources and rhythms.

Chapters 6 and 7 develop the key point B. Chapter 6 extends the framework introduced in Chapter 3, demonstrating its applicability with only 10 minutes of recording time, which is an important step toward research in particular cohorts. In addition, Chapter 7 demonstrates that this analysis approach is not only applicable, but also effective in studying speech and language impairment. Specifically, measures of phonological processing were affected by dyslexia, which is in line with theories stating that this reading disorder is actually underpinned by a phonological deficit.

The rest of this chapter focuses on more general discussion that relates to the thesis as whole. The primary aims of the thesis are dealt with in Sections 8.1, 8.2, and 8.3. Limitations and (ongoing and) future applications of the work are discussed in Section 8.4, followed by concluding remarks.

8.1 A novel approach to study natural speech perception

This thesis finds its roots in the relatively recent discovery that low-frequency cortical activity tracks the temporal envelope of natural speech (Ahissar et al., 2001; Luo and Poeppel, 2007; Aiken and Picton, 2008). This has led to insights into core mechanisms of speech perception, such as selective attention (Ding and Simon, 2012a; Power et al., 2012; O'Sullivan et al., 2014) and audio-visual integration (Luo et al., 2010; Zion Golumbic et al., 2013a; Crosse et al., 2015; Crosse et al., 2016a). This work allows the study of these mechanisms in naturalistic conditions, however this research has been limited in one important respect. Namely, it has been unclear whether envelope tracking reflects any speech-specific cortical processing or whether it simply indexes passive following of the acoustic energy of any sound input. The first main finding of this thesis answers this question by demonstrating that EEG-measured low-frequency cortical oscillations entrain to phonological features of speech.

This discovery was achieved through the design of an analysis framework based on so-called forward models, which predict the EEG signal using different representations of the speech. The quantitative results have been suggested to reflect the amount of speech information encoded in the neural signal, which have been related specifically to acoustic (E, S) and phonetic features (P, F) of speech. The fact that the combined model FS produced higher EEG predictions than the purely acoustic and phonetic ones, and that this was not the case when participants were presented with time-reversed speech, has been interpreted as a proof that acoustic and phonetic encoding in the EEG were not completely overlapping and that, therefore, EEG reflects phonetic encoding.

The motivation behind the use of time-reversed speech in the control experiment was to have a stimulus with similar properties to the original audio-book but that was unintelligible. Of course, forward and time-reversed speech share the same spectro-temporal properties with the sole difference that one is the time inversion of the other, which allowed for a fair comparison. One other option that we considered was to use different languages (not spoken by the participants) with similar phonological properties to English, such as Dutch, and with different phonological inventory, such as Hindi. However, time-reversed speech was a better choice for a fair control experiment for two main reasons: Firstly, a foreign language may have various differences with English (e.g., spectro-temporal, rhythmic, phonological), which would make the comparison between natural speech and control conditions depending not only on comprehension, but also on such lower-level differences. Secondly, the use of a different automatic phoneme aligner software would be necessary, which would require a careful comparison of the alignment precision between English and the foreign language (usually the most common languages have the best aligners). For these reasons, time-reversed speech was an optimal choice for that specific control experiment. Of course, this transformation entails effects such as reduced attention, because the stimulus becomes unintelligible, and changes of the temporal dynamics at the onsets and offsets (of words and phonemes). Chapter 4 provides further support to both validity and effectiveness of our framework. In fact, differently from Chapter 3, perceived clarity was modulated without changing the speech acoustics and that change correlated with our novel neural index FS-S.

This framework provides an unprecedented opportunity to characterise the acoustic-phonetic processing of natural speech with high spatial and temporal resolution using a macroscopic technique that can capture activity in auditory cortex as well as concurrent activity in frontal and visual cortices. Importantly, this finding has been

reproduced by subsequent research. Firstly, the work in Chapter 4 confirmed the results for the natural speech condition using a different experimental paradigm. This study indicated that EEG predictability increases with spectral detail (Section 4.3.2) and is modulated by prior knowledge. Importantly, the measure FS-S increased with perceived intelligibility, which demonstrates that this index is speech-specific. Also, recent work from a different group provided further support to our findings by investigating phonological perception on continuous speech using an ERP analysis (Khalighinejad et al., 2017). It was observed that EEG responses to continuous speech encode both phonetic and speaker information at various latencies relative the phoneme onset. Differently from linear regression approaches, event-related averaging has the drawback of not addressing explicitly the issue of overlapping between long and short latency responses to consecutive stimuli. Despite these limitations, Khalighinejad et al. reported remarkably distinct cortical responses that are in line with the finding discussed in this thesis. Nevertheless, the choice of using the mTRF approach enabled us to further develop the work from Chapter 3 by using EEG predictability and generic models, which could be difficult with event-related averaging which, differently from regression approaches, is not optimised for predictions.

8.2 Heterogeneous roles of low-frequency cortical rhythms

Recent literature on speech perception hypothesised that processing of speech units at specific rates, such as phonemes, engages the two hemispheres differently (Poeppel, 2003). Although this has been shown in a variety of experiments (Giraud et al., 2007; Hickok and Poeppel, 2007; Gross et al., 2013; Keitel et al., 2017), there remains significant controversy on the specific roles of different cortical rhythms. One view is that delta- and high-frequency activity (> 40 Hz) are reliable indicators of perceived linguistic representations, while theta-band entrainment may primarily reflect the analysis of acoustic features of speech (Kösem and van Wassenhove, 2016). However, as mentioned in Chapter 4, this interpretation does not agree with previous (tentative) suggestions that speech features critical for intelligibility are reflected by theta-band oscillations, while delta-band entrainment relates to the acoustic rhythm (Ding and Simon, 2014).

The greater specificity afforded by our forward encoding models provides insights from a different perspective from that of previous studies based on broader envelope

entrainment measures. Our considerations focus on delta- and theta-band neural rhythms because EEG measures of cortical entrainment were mostly driven by those frequencies. In particular, while Chapter 3 demonstrates that both EEG delta- and theta-bands reflect phoneme-level entrainment, a major distinction between purely acoustic and phonetic models emerged in the slower delta rhythms. This suggests that theta-band more strongly reflects entrainment to speech acoustics. This finding is confirmed in Chapter 4, where only delta-band FS-S increased with perceived intelligibility, which confirms the speech-specific role of these EEG rhythms. Also, Chapter 5 showed a distinct role of delta and theta oscillations in envelope entrainment. While theta-band showed "only" a suppression at short latencies, delta-band revealed spatio-temporally complex patterns due to prior knowledge and increased intelligibility that involved latencies up to 400 ms from stimulus onset. Unfortunately, the experimental paradigm, originally developed for a cross-correlation approach, could not be used for phoneme-level analysis because of the limited amount of stimulus data (only 2 short sentences were repeated throughout the experiment). Phoneme-level analysis was conducted instead in Chapter 7, where results were driven by a combination of delta- and theta-bands with, once again, a major contribution from EEG delta-band.

These results support the distinctive complementary roles of the two bands in the processing of acoustic and phonological features. Such roles may extend over the temporal lobe and involve frontal regions such as the IFG. Of course, this thesis focused on a rigid separation of delta- and theta-bands as 1-4 Hz and 4-8 Hz, which may be considered standard (as for other variations, such as 1-3 Hz and 3-7 Hz). Further work will have to verify whether there is an actual "hard" boundary between these cortical oscillations (or, at least, that we can model it as hard), or if a "softer" boundary that adapts to the specific speech rate of a stimulus should be applied.

8.3 The role of prior predictive knowledge in speech comprehension

Recent studies have contributed to the characterisation of specific cortical areas in terms of their functional roles in speech comprehension. In particular, a hierarchical organisation of temporal areas supporting the perceptual and lexical processing of speech has been identified: Key regions include the superior temporal gyrus (Humphries et al., 2014) and the superior temporal sulcus (Chang et al., 2010; Mesgarani et al., 2014;

Overath et al., 2015), which exhibited sensitivity to acoustic and phonetic features of speech. Furthermore, the middle temporal gyrus has been related to higher-level lexical processing (Lau et al., 2008; Turken and Dronkers, 2011).

Although these studies provide important insights into the mechanisms of key regions in the speech network, a number of fundamental questions remain unanswered, especially regarding the temporal dynamics and interactions between cortical areas. The presence of both feed-forward and feed-back connections allows for the efficient and effortless integration of sensory auditory input with prior knowledge. These mechanisms have been shown to extend over the temporal lobe. In particular, prior knowledge has been shown to elicit a top-down influence within temporal areas (Tuennerhoff and Noppeney, 2016) and from inferior frontal gyrus to STG (Sohoglu et al., 2012; Leonard et al., 2016; Sohoglu and Davis, 2016).

Theories on more general neural mechanisms, such as sharpening and predictive coding, may explain these effects (Section 2.2.2). As known from Chapter 2, these frameworks predict various patterns of suppression and enhancement of neural activity with prior knowledge. Despite recent tantalising discussions on this topic, which touch on various phenomenon and diseases (from speech perception to visions in schizophrenia), scientists have not reached an agreement (Clark, 2013). In the context of speech perception, a recent ECoG study from Holdgraf et al. (2016) demonstrated that prior knowledge, rather than simply producing increases or decreases in activation in specific cortical areas, enhances high gamma-band entrainment to speech-like spectro-temporal features in temporal cortex. In this sense, this study agrees with another recent finding from Blank and Davis (2016) that showed a double effect of prior knowledge in STS: a suppression in fMRI BOLD response and an enhanced representation of speech-specific features. While Blank and Davis explicitly linked their finding with the predictive coding theory, the results reported by Holdgraf et al. appear to be more in line with the sharpening theory, as they showed enhancement but not a suppression of cortical responses to speech.

The results from Chapters 4 and 5 provide new important insights that complement the current literature. Firstly, Chapter 4 demonstrated a dual effect of prior knowledge: an increase with perceived intelligibility and a suppression. Both enhancement and suppression effects also emerged in Chapter 5, where they were specific to delta- and theta-bands respectively. These results are in line with the finding from Blank and Davis. In particular, we contend that there is a link between the delta-band phoneme-level

increase (FS-S) in Chapter 4, the delta-band envelope entrainment enhancement measured for STS in Chapter 5, and the enhancement of speech-specific information encoding in STS described by Blank and Davis. Indeed, a review of predictive coding theory has proposed that there may exist two distinct units within our sensory processing hierarchies: representational units and error units (Hohwy, 2013). And this idea fits well with our dual effects. It may be the case that activity from representational units is increased with prior knowledge in our experiment, while activity from error units is suppressed. This account may address also the result from Holdgraf and colleagues. In fact, it is possible that the ECoG measures and cortical locations used in that study were sensitive only (or mainly) to representational units, which would explain why no suppression emerged from that study.

Within the same theoretical framework, there are other (complementary) explanations for a dual effect that includes both enhancement and suppression due to prior knowledge. One of these suggests that higher-order areas that are "inactive" or less responsive when speech is less (or not) intelligible would become more active when prior knowledge makes speech clearer. Such a phenomenon would explain the results in Chapter 5, where a causal bottom-up link from STS to MTG emerges when prior knowledge is available, suggesting that MTG is involved in the processing of speech features that require intelligibility.

Indeed, the results presented in this thesis indicate that the complex underpinnings of speech perception cannot be captured by 2- or 3-node models as previously attempted (Sohoglu and Davis, 2016; Tuennerhoff and Noppeney, 2016). In particular, the study in Chapter 5 suggests that IFG may interact differently with distinct temporal areas, which could not be captured by a two-node model. In addition, the same chapter indicates a need for high temporal and spatial resolution to disentangle feed-forward and feed-back effects.

The results in this thesis provide an opportunity to discuss theories that go beyond speech perception and the auditory modality. Overall, the present work supports the predictive coding framework and indicates that such enhancement/suppressive effects apply to speech processing at the phonemic level. Furthermore, a cortical hierarchy that involves fast feedback effects (from IFG to HG) and slower feed-forward processing (from HG, STS, to MTG) emerged as part of the mechanism that integrates prior knowledge with sensory input and producing a perceptual enhancement.

8.4 Ongoing and Future Work

"The actual words are the tip of a vast iceberg of very rapid unconscious non-linguistic processing" (Prof. Steven Pinker)

The categorical identification of phonemes constitutes a first step toward the abstraction of a sensory input, which becomes progressively less modality-specific and, therefore, more abstract. For instance, my concept of 'apple' is an abstraction of what that word represents for me. It could be taste, visual, or even auditory (e.g., the sound while eating one or the pronunciation of the actual word 'apple'). Whatever the modality, this internal representation may constitute one important difference between people's perceptual experiences and the ability to study these abstractions may involve a variety of fields such as cognitive neuroscience, clinical and criminal psychology, computer engineering, marketing, and even religion. This thesis may be nothing of all of that, or it may represent a first step toward the ability of modelling this abstract internal representation of our daily perceptual experience.

The research presented in this manuscript demonstrated that non-invasive EEG reflects the cortical processing of natural speech at the level of phonemes. A novel analysis framework was developed to derive quantitative objective measures of neural entrainment to acoustic and phonological features of speech. Indeed, the cortical processing of speech requires the brain to decode information at higher levels of abstraction than phonemes, e.g., grammar and semantics. Further research is needed to model and quantify cortical responses to such abstract features of speech.

One ongoing line of research that we are currently pursuing is the study of the visual component of speech. In particular, one core question is whether categorical perception of the so-called visemes – i.e. the visual equivalent of phonemes – occur in visual cortex during lip-reading. Using a similar approach to the one in Chapter 3, some results have been obtained in this direction, suggesting some level of categorical perception in visual areas during silent lip-reading (O'Sullivan et al., 2016). However, this work is far from a conclusive demonstration of such a phenomenon and further research is needed.

A second line of research that we have been investigating is the extension of this analysis framework to other linguistic components. In particular, we have been investigating the possibility of modelling EEG responses to speech based on the semantic content within each sentence (Broderick, In preparation). We envisage that the ability to do this would constitute a significant advance of the current analysis methodologies and

that it has the potential to provide new insights on linguistic processing. Crucially, these and other possible measures of linguistic processing could be extracted from the same recordings used for acoustics and phonetics. Potentially, one unique natural speech experiment, as in Chapter 3, 6, and 7, could produce a whole set of neural measures at different linguistics levels, while a long battery of various tests is currently required by standardised tests to extract behavioural measures of language skills. We contend that the success of this research may lead to a new, more direct and efficient way to study speech perception and, possibly, to conduct diagnosis of speech and language impairments.

Of course, much more work is needed to clarify whether this research can be successful. Chapters 6 and 7 constitutes "only" the first step toward such objectives. Firstly, the effectiveness of our novel analysis approach has to be tested on populations with other language deficits. Secondly, the approach may be further strengthened with the use of mapping approaches more sophisticated than a regularised linear regression. In addition, the feasibility of this approach using EEG systems with dry electrodes should be verified. In fact, this technology allows for a shorter set-up time, which would be preferable in many cases, including longitudinal studies, which are likely to involve many sessions, and specific populations that cannot commit to an experiment for a long time, e.g., elderly with dementia.

As a concluding note, this research work can be beneficial to better understand the cortical mechanisms that underpin speech processing, and has already provided novel insights into such neural functions. In addition to what previously said, this work may provide a way to investigate speech processing on populations where standard behavioural methods are not reliable or absent. One example is the study of the development of speech processing, starting from new-borns. In particular, the ability to derive reliable objective measures of speech processing in infants and young children may have the potential to allow early diagnosis of speech and language impairment, which currently can be very hard and may require the awaiting for more severe symptoms to arise before confidently achieve a diagnosis. Furthermore, this approach could reveal new insights on other important and not well understood phenomena, such as second language acquisition and bilingualism.

8.5 Summary and Conclusions

The body of research from this thesis provides a method to investigate the cortical processing of natural speech with non-invasive EEG and to disambiguate cortical responses to speech acoustics and phonetics. This methodology has been used to investigate the cortical mechanisms that underpin the integration of prior knowledge with sensory input, revealing new insights that support the theory of predictive coding and the interaction between frontal and temporal areas during such process. Furthermore, this thesis demonstrated the applicability of our novel analysis approach using less than 10 minutes of EEG recordings, during which participants listened to a continuous audio-story, and its effectiveness was demonstrated by studying speech processing in children with dyslexia.

Bibliography

Abrams DA, Nicol T, Zecker S, Kraus N (2009) Abnormal cortical processing of the syllable rate of speech in poor readers. The Journal of neuroscience : the official journal of the Society for Neuroscience 29:7686-7693.

Ahissar E, Nagarajan S, Ahissar M, Protopapas A, Mahncke H, Merzenich MM (2001) Speech Comprehension is Correlated with Temporal Response Patterns Recorded from Auditory Cortex. Proceedings of the National Academy of Sciences of the United States of America 98:13367-13372.

Ahissar M, Hochstein S (2004) The reverse hierarchy theory of visual perceptual learning. Trends Cogn Sci 8:457-464.

Ahlfors SP, Han J, Belliveau JW, Hämäläinen MS (2010) Sensitivity of MEG and EEG to Source Orientation. Brain Topography 23:227-232.

Aiken SJ, Picton TW (2008) Human cortical responses to the speech envelope. Ear Hear 29:139-157.

Alink A, Schwiedrzik CM, Kohler A, Singer W, Muckli L (2010) Stimulus predictability reduces responses in primary visual cortex. J Neurosci 30:2960-2966.

American Speech-Language-Hearing Association (1993) Definitions of communication disorders and variations. Available from www.asha.org/policy.

Arnal LH, Giraud AL (2012) Cortical oscillations and sensory predictions. Trends Cogn Sci 16:390-398.

Arnal LH, Wyart V, Giraud A-L (2011) Transitions in neural oscillations reflect prediction errors generated in audiovisual speech. Nat Neurosci 14:797-801.

Arnal LH, Doelling KB, Poeppel D (2015) Delta–Beta Coupled Oscillations Underlie Temporal Prediction Accuracy. Cerebral Cortex (New York, NY) 25:3077-3085.

Aslin RN, Mehler J (2005) Near-infrared spectroscopy for functional studies of brain activity in human infants: promise, prospects, and challenges. Journal of Biomedical Optics 10:011009-0110093.

Aydin Ü, Vorwerk J, Dümpelmann M, Küpper P, Kugel H, Heers M, Wellmer J, Kellinghaus C, Haueisen J, Rampp S, Stefan H, Wolters CH (2015) Combined EEG/MEG Can Outperform Single Modality EEG or MEG Source Reconstruction in Presurgical Epilepsy Diagnosis. PLOS ONE 10:e0118753.

Baddeley A (2003) Working memory and language: An overview. Journal of communication disorders 36:189-208.

Baillet S (2017) Magnetoencephalography for brain electrophysiology and imaging. Nat Neurosci 20:327-339.

Baillet S, Garnero L, Marin G, Hugonin JP (1999) Combined MEG and EEG source imaging by minimization of mutual information. IEEE Trans Biomed Eng 46:522-534.
Baker SF, Ireland JL (2007) The link between dyslexic traits, executive functioning, impulsivity and social self-esteem among an offender and non-offender sample. International journal of law and psychiatry 30:492-503.
Baltzell LS, Horton C, Shen Y, Richards VM, D'Zmura M, Srinivasan R (2016) Attention selectively modulates cortical entrainment in different regions of the speech spectrum. Brain Res 1644:203-212.
Bar M, Kassam KS, Ghuman AS, Boshyan J, Schmid AM, Dale AM, Hämäläinen MS, Marinkovic K, Schacter DL, Rosen BR, Halgren E (2006) Top-down facilitation of visual recognition. Proceedings of the National Academy of Sciences of the United States of America 103:449-454.
Baumann S, Petkov CI, Griffiths TD (2013) A unified framework for the organization of the primate auditory cortex. Frontiers in systems neuroscience 7:11.
Benjamini Y, Yekutieli D (2001) The Control of the False Discovery Rate in Multiple Testing under Dependency. The Annals of Statistics 29:1165-1188.
Bialek W, Rieke F, Vansteveninck RRD, Warland D (1991) READING A NEURAL CODE. Science 252:1854-1857.
Binder JR, Frost JA, Hammeke TA, Bellgowan PS, Springer JA, Kaufman JN, Possing ET (2000) Human temporal lobe activation by speech and nonspeech sounds. Cerebral cortex (New York, NY : 1991) 10:512-528.
Bishop D (2003a) Test for Reception of Grammar. Version 2. In: London, Stockholm, The Psychological Corporation. Harcourt Assessment.
Bishop DV (2003b) The Children's Communication Checklist: CCC-2: ASHA.
Blank H, Davis MH (2016) Prediction Errors but Not Sharpened Signals Simulate Multivoxel fMRI Patterns during Speech Perception. PLoS Biol 14:e1002577.
Boemio A, Fromm S, Braun A, Poeppel D (2005) Hierarchical and asymmetric temporal sensitivity in human auditory cortices. Nat Neurosci 8:389-395.
Bonte M, Parviainen T, Hytönen K, Salmelin R (2006) Time course of top-down and bottom-up influences on syllable processing in the auditory cortex. Cerebral Cortex 16:115-123.
Bornkessel-Schlesewsky I, Schlesewsky M, Small SL, Rauschecker JP (2015) Neurobiological roots of language in primate audition: common computational properties. Trends in cognitive sciences 19:142-150.
Broderick MA, Anderson AJ, Di Liberto GM, Lalor EC (In preparation) Electrophysiological responses to natural, ongoing speech encode a measure of semantic surprisal.
Brodmann K (1909) Vergleichende Lokalisationslehre der Grosshirnrinde in ihren Prinzipien dargestellt auf Grund des Zellenbaues: Johann Ambrosius Barth Verlag.
Brooks MY (2014) School, Disability Status, and Delinquency: An Examination of Delinquency Among Rural Adolescents.
Büchel C, Geuter S, Sprenger C, Eippert F (2014) Placebo Analgesia: A Predictive Coding Perspective. Neuron 81:1223-1239.
Buzsáki G, Anastassiou CA, Koch C (2012) The origin of extracellular fields and currents — EEG, ECoG, LFP and spikes. Nature reviews Neuroscience 13:407-420.
Casserly ED, Pisoni DB (2010) Speech perception and production. Wiley interdisciplinary reviews Cognitive science 1:629-647.

Catts HW, Fey ME, Tomblin JB, Zhang X (2002) A longitudinal investigation of reading outcomes in children with language impairments. Journal of speech, language, and hearing research : JSLHR 45:1142-1157.
Celesia GG (1976) Organization of auditory cortical areas in man. Brain 99:403-414.
Chait M, Greenberg S, Arai T, Simon JZ, Poeppel D (2015) Multi-time resolution analysis of speech: evidence from psychophysics. Front Neurosci 9:214.
Chang EF, Rieger JW, Johnson K, Berger MS, Barbaro NM, Knight RT (2010) Categorical speech representation in human superior temporal gyrus. Nat Neurosci 13:1428-1432.
Chennu S, Noreika V, Gueorguiev D, Blenkmann A, Kochen S, Ibáñez A, Owen AM, Bekinschtein TA (2013) Expectation and Attention in Hierarchical Auditory Prediction. The Journal of Neuroscience 33:11194-11205.
Chinchor N (1992) MUC-4 evaluation metrics. In: Proceedings of the 4th conference on Message understanding, pp 22-29. McLean, Virginia: Association for Computational Linguistics.
Chittka L, Brockmann A (2005) Perception Space—The Final Frontier. PLoS Biol 3:e137.
Chomsky N, Halle M (1968) The sound pattern of English.
Chomsky N, Miller GA (1968) Introduction to the formal analysis of natural languages.
Clark A (2013) Whatever next? Predictive brains, situated agents, and the future of cognitive science. The Behavioral and brain sciences 36:181-204.
Clark A (2016) Surfing Uncertainty: Prediction, Action, and the Embodied Mind: Oxford University Press.
Clark G (2006) Cochlear Implants: Fundamentals and Applications: Springer New York.
Clark KA, Helland T, Specht K, Narr KL, Manis FR, Toga AW, Hugdahl K (2014) Neuroanatomical precursors of dyslexia identified from pre-reading through to age 11. Brain 137:3136-3141.
Clark MKK, A. G. (2010) Language Disorders (Child Language Disorders). JH Stone, M Blouin, editors International Encyclopedia of Rehabilitation.
Clegg J, Henderson J (1999) Developmental language disorders: changing economic costs from childhood into adult life. Mental Health Research Review 6:27-30.
Cogan GB, Thesen T, Carlson C, Doyle W, Devinsky O, Pesaran B (2014) Sensory-motor transformations for speech occur bilaterally. Nature 507:94-98.
Cohen D, Cuffin BN (1987) A method for combining MEG and EEG to determine the sources. Phys Med Biol 32:85-89.
Cone-Wesson B, Wunderlich J (2003) Auditory evoked potentials from the cortex: audiology applications. Curr Opin Otolaryngol Head Neck Surg 11:372-377.
Connine CM, Clifton C, Jr. (1987) Interactive use of lexical information in speech perception. Journal of experimental psychology Human perception and performance 13:291-299.
Corby JC, Kopell BS (1972) Differential contributions of blinks and vertical eye movements as artifacts in EEG recording. Psychophysiology 9:640-644.
Corey DM, Dunlap WP, Burke MJ (1998) Averaging correlations: Expected values and bias in combined Pearson rs and Fisher's z transformations. The Journal of general psychology 125:245-261.
Crosse MJ, Butler JS, Lalor EC (2015) Congruent visual speech enhances cortical entrainment to continuous auditory speech in noise-free conditions. The Journal of Neuroscience 35:14195–11420.
Crosse MJ, Di Liberto GM, Lalor EC (2016a) Eye Can Hear Clearly Now: Inverse Effectiveness in Natural Audiovisual Speech Processing Relies on Long-Term

Crossmodal Temporal Integration. The Journal of neuroscience : the official journal of the Society for Neuroscience 36:9888-9895.
Crosse MJ, Di Liberto GM, Bednar A, Lalor EC (2016b) The Multivariate Temporal Response Function (mTRF) Toolbox: A MATLAB Toolbox for Relating Neural Signals to Continuous Stimuli. Frontiers in Human Neuroscience 10.
Damasio H, Damasio AR (1980) The anatomical basis of conduction aphasia. Brain 103:337-350.
Daniel SS, Walsh AK, Goldston DB, Arnold EM, Reboussin BA, Wood FB (2006) Suicidality, school dropout, and reading problems among adolescents. Journal of learning disabilities 39:507-514.
Davis MH, Johnsrude IS (2003) Hierarchical processing in spoken language comprehension. The Journal of Neuroscience 23:3423-3431.
Davis MH, Johnsrude IS (2007) Hearing speech sounds: top-down influences on the interface between audition and speech perception. Hear Res 229:132-147.
Davis MH, Johnsrude IS, Hervais-Adelman A, Taylor K, McGettigan C (2005) Lexical information drives perceptual learning of distorted speech: evidence from the comprehension of noise-vocoded sentences. Journal of Experimental Psychology: General 134:222.
de Boer R, Kuyper P (1968) Triggered correlation. IEEE Trans Biomed Eng 15:169-179.
Delorme A, Makeig S (2004) EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. Journal of neuroscience methods 134:9-21.
Delorme A, Mullen T, Kothe C, Akalin Acar Z, Bigdely-Shamlo N, Vankov A, Makeig S (2011) EEGLAB, SIFT, NFT, BCILAB, and ERICA: New Tools for Advanced EEG Processing. Computational Intelligence and Neuroscience 2011:12.
den Ouden HE, Friston KJ, Daw ND, McIntosh AR, Stephan KE (2009) A dual role for prediction error in associative learning. Cerebral cortex (New York, NY : 1991) 19:1175-1185.
Depireux DA, Simon JZ, Klein DJ, Shamma SA (2001) Spectro-Temporal Response Field Characterization With Dynamic Ripples in Ferret Primary Auditory Cortex. J Neurophysiol 85:1220-1234.
DeWitt I, Rauschecker JP (2012) Phoneme and word recognition in the auditory ventral stream. Proceedings of the National Academy of Sciences of the United States of America 109:E505-514.
Di Liberto GM, Lalor EC (2017) Indexing cortical entrainment to natural speech at the phonemic level: Methodological considerations for applied research. Hearing research 348:70-77.
Di Liberto GM, O'Sullivan JA, Lalor EC (2015) Low-Frequency Cortical Entrainment to Speech Reflects Phoneme-Level Processing. Curr Biol 25:2457-2465.
Di Liberto GM, Crosse MJ, Lalor EC (in review) Cortical measures of phoneme-level speech encoding correlate with the perceived clarity of natural speech. JN-RM-0648-17.
Ding M, Chen Y, Bressler SL (2006) 17 Granger Causality: Basic Theory and Application to Neuroscience. Handbook of time series analysis: recent theoretical developments and applications:437.
Ding M, Bressler SL, Yang W, Liang H (2000) Short-window spectral analysis of cortical event-related potentials by adaptive multivariate autoregressive modeling: data preprocessing, model validation, and variability assessment. Biological Cybernetics 83:35-45.

Ding N, Simon JZ (2012a) Emergence of neural encoding of auditory objects while listening to competing speakers. Proceedings of the National Academy of Sciences of the United States of America 109:11854-11859.
Ding N, Simon JZ (2012b) Neural coding of continuous speech in auditory cortex during monaural and dichotic listening. J Neurophysiol 107:78-89.
Ding N, Simon JZ (2013) Adaptive temporal encoding leads to a background-insensitive cortical representation of speech. The Journal of neuroscience : the official journal of the Society for Neuroscience 33:5728-5735.
Ding N, Simon JZ (2014) Cortical entrainment to continuous speech: functional roles and interpretations. Front Hum Neurosci 8:311.
Ding N, Chatterjee M, Simon JZ (2014) Robust cortical entrainment to the speech envelope relies on the spectro-temporal fine structure. NeuroImage 88:41-46.
Dudley H (1939) Remaking Speech. The Journal of the Acoustical Society of America 11:169-177.
Egner T, Monti JM, Summerfield C (2010) Expectation and surprise determine neural population responses in the ventral visual stream. J Neurosci 30:16601-16608.
Ehret G, Romand R (1997) The Central Auditory System: Oxford University Press.
Ernst MO, Banks MS (2002) Humans integrate visual and haptic information in a statistically optimal fashion. Nature 415:429-433.
Evans AC, Collins DL, Mills SR, Brown ED, Kelly RL, Peters TM (1993) 3D statistical neuroanatomical models from 305 MRI volumes. In: 1993 IEEE Conference Record Nuclear Science Symposium and Medical Imaging Conference, pp 1813-1817 vol.1813.
Flanagan DP, Genshaft J, Harrison PL (1997) Contemporary Intellectual Assessment: Theories, Tests, and Issues: Guilford Press.
Flinker A, Knight Robert T (2016) A Cool Approach to Probing Speech Cortex. Neuron 89:1123-1125.
Fontolan L, Morillon B, Liegeois-Chauvel C, Giraud AL (2014) The contribution of frequency-specific activity to hierarchical information processing in the human auditory cortex. Nat Commun 5:4694.
Ford L, Dahinten VS (2005) Use of intelligence tests in the assessment of preschoolers. Contemporary intellectual assessment:487-503.
Foster JR, Summerfield AQ, Marshall DH, Palmer L, Ball V, Rosen S (1993) Lip-reading the BKB sentence lists: Corrections for list and practice effects. British Journal of Audiology 27:233-246.
Freeman WJ, Holmes MD, Burke BC, Vanhatalo S (2003) Spatial spectra of scalp EEG and EMG from awake humans. Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology 114:1053-1068.
Friederici AD (2011) The Brain Basis of Language Processing: From Structure to Function. Physiological Reviews 91:1357-1392.
Friederici AD (2012) The cortical language circuit: from auditory perception to sentence comprehension. Trends Cogn Sci 16:262-268.
Friston K (2005) A theory of cortical responses. Philosophical transactions of the Royal Society of London Series B, Biological sciences 360:815-836.
Friston K, Kiebel S (2009) Predictive coding under the free-energy principle. Philosophical transactions of the Royal Society of London Series B, Biological sciences 364:1211-1221.
Fuchs M, Wagner M, Wischmann HA, Kohler T, Theissen A, Drenckhahn R, Buchner H (1998) Improving source reconstructions by combining bioelectric and

biomagnetic data. Electroencephalography and clinical neurophysiology 107:93-111.
Galaburda A, Sanides F (1980) Cytoarchitectonic organization of the human auditory cortex. The Journal of comparative neurology 190:597-610.
Ganong WF, 3rd (1980) Phonetic categorization in auditory word perception. Journal of experimental psychology Human perception and performance 6:110-125.
Gardner H, Froud K, McClelland A, van der Lely HK (2006) Development of the Grammar and Phonology Screening (GAPS) test to assess key markers of specific language and literacy difficulties in young children. International Journal of Language & Communication Disorders 41:513-540.
Gardner MPH, Fontanini A (2014) Encoding and Tracking of Outcome-Specific Expectancy in the Gustatory Cortex of Alert Rats. The Journal of Neuroscience 34:13000-13017.
Gazzaniga MS (2009) The Cognitive Neurosciences: MIT Press.
George N, Dolan RJ, Fink GR, Baylis GC, Russell C, Driver J (1999) Contrast polarity and face recognition in the human fusiform gyrus. Nat Neurosci 2:574-580.
Gervais H, Belin P, Boddaert N, Leboyer M, Coez A, Sfaello I, Barthelemy C, Brunelle F, Samson Y, Zilbovicius M (2004) Abnormal cortical voice processing in autism. Nat Neurosci 7:801-802.
Giraud AL, Poeppel D (2012) Cortical oscillations and speech processing: emerging computational principles and operations. Nat Neurosci 15:511-517.
Giraud AL, Kleinschmidt A, Poeppel D, Lund TE, Frackowiak RS, Laufs H (2007) Endogenous cortical rhythms determine cerebral specialization for speech perception and production. Neuron 56:1127-1134.
Glover GH (2011) Overview of Functional Magnetic Resonance Imaging. Neurosurgery clinics of North America 22:133-139.
Goldstone RL (1998) Perceptual learning. Annual review of psychology 49:585-612.
Goncharova, II, McFarland DJ, Vaughan TM, Wolpaw JR (2003) EMG contamination of EEG: spectral and topographical characteristics. Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology 114:1580-1593.
Gorman K, Howell J, Wagner M (2011) Prosodylab-aligner: A tool for forced alignment of laboratory speech. 2011 39:2.
Goswami U (2011) A temporal sampling framework for developmental dyslexia. Trends Cogn Sci 15:3-10.
Goswami U (2015) Sensory theories of developmental dyslexia: three challenges for research. Nature reviews Neuroscience 16:43-54.
Goswami U, Leong V (2013) Speech rhythm and temporal structure: converging perspectives. Lab Phonol 4:67-92.
Granger CWJ (1969) Investigating Causal Relations by Econometric Models and Cross-spectral Methods. Econometrica 37:424-438.
Greenwood DD (1961) Auditory Masking and the Critical Band. The Journal of the Acoustical Society of America 33:484-502.
Grill-Spector K, Henson R, Martin A (2006) Repetition and the brain: neural models of stimulus-specific effects. Trends in Cognitive Sciences 10:14-23.
Gross J, Hoogenboom N, Thut G, Schyns P, Panzeri S, Belin P, Garrod S (2013) Speech rhythms and multiplexed oscillatory sensory coding in the human brain. PLoS Biol 11:e1001752.

Hackett TA, Stepniewska I, Kaas JH (1998) Subdivisions of auditory cortex and ipsilateral cortical connections of the parabelt auditory cortex in macaque monkeys. The Journal of comparative neurology 394:475-495.
Hackett TA, Preuss TM, Kaas JH (2001) Architectonic identification of the core region in auditory cortex of macaques, chimpanzees, and humans. The Journal of comparative neurology 441:197-222.
Hagoort P (2005) On Broca, brain, and binding: a new framework. Trends in Cognitive Sciences 9:416-423.
Hamalainen JA, Rupp A, Soltesz F, Szucs D, Goswami U (2012) Reduced phase locking to slow amplitude modulation in adults with dyslexia: an MEG study. Neuroimage 59:2952-2961.
Handjaras G, Ricciardi E, Leo A, Lenci A, Cecchetti L, Cosottini M, Marotta G, Pietrini P (2016) How concepts are encoded in the human brain: A modality independent, category-based cortical organization of semantic knowledge. NeuroImage 135:232-242.
Happe F, Ronald A, Plomin R (2006) Time to give up on a single explanation for autism. Nat Neurosci 9:1218-1220.
Heald SL, Nusbaum HC (2014) Speech perception as an active cognitive process. Frontiers in systems neuroscience 8:35.
Heim S, Eulitz C, Elbert T (2003) Altered hemispheric asymmetry of auditory P100m in dyslexia. Eur J Neurosci 17:1715-1722.
Hein G, Knight RT (2008) Superior temporal sulcus--It's my area: or is it? Journal of cognitive neuroscience 20:2125-2136.
Henseler I, Mädebach A, Kotz SA, Jescheniak JD (2014) Modulating Brain Mechanisms Resolving Lexico-semantic Interference during Word Production: A Transcranial Direct Current Stimulation Study. Journal of cognitive neuroscience 26:1403-1417.
Herculano-Houzel S (2009) The Human Brain in Numbers: A Linearly Scaled-up Primate Brain. Frontiers in Human Neuroscience 3:31.
Hickok G, Poeppel D (2004) Dorsal and ventral streams: a framework for understanding aspects of the functional anatomy of language. Cognition 92:67-99.
Hickok G, Poeppel D (2007) The cortical organization of speech processing. Nature reviews Neuroscience 8:393-402.
Hickok G, Small SL (2015) Neurobiology of Language: Elsevier Science.
Hohwy J (2013) The predictive mind: Oxford University Press.
Holdgraf CR, de Heer W, Pasley B, Rieger J, Crone N, Lin JJ, Knight RT, Theunissen FE (2016) Rapid tuning shifts in human auditory cortex enhance speech intelligibility. Nat Commun 7:13654.
Howard MF, Poeppel D (2010) Discrimination of Speech Stimuli Based on Neuronal Response Phase Patterns Depends on Acoustics But Not Comprehension. J Neurophysiol 104:2500-2511.
Hsieh P-J, Vul E, Kanwisher N (2010) Recognition Alters the Spatial Pattern of fMRI Activation in Early Retinotopic Cortex. J Neurophysiol 103:1501-1507.
Huang MX, Mosher JC, Leahy RM (1999) A sensor-weighted overlapping-sphere head model and exhaustive head model comparison for MEG. Phys Med Biol 44:423-440.
Huang MX, Shih JJ, Lee RR, Harrington DL, Thoma RJ, Weisend MP, Hanlon F, Paulson KM, Li T, Martin K, Millers GA, Canive JM (2004) Commonalities and differences among vectorized beamformers in electromagnetic source imaging. Brain Topogr 16:139-158.

Huang MX, Edgar JC, Thoma RJ, Hanlon FM, Moses SN, Lee RR, Paulson KM, Weisend MP, Irwin JG, Bustillo JR, Adler LE, Miller GA, Canive JM (2003) Predicting EEG responses using MEG sources in superior temporal gyrus reveals source asynchrony in patients with schizophrenia. Clinical Neurophysiology 114:835-850.
Humphries C, Sabri M, Lewis K, Liebenthal E (2014) Hierarchical organization of speech perception in human auditory cortex. Front Neurosci 8.
Iverson P, Kuhl PK (1995) Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling. J Acoust Soc Am 97:553-562.
Jackson AF, Bolger DJ (2014) The neurophysiological bases of EEG and EEG measurement: A review for the rest of us. Psychophysiology 51:1061-1071.
Jakobson R, Fant G, Halle M (1969) Preliminaries to Speech Analysis: The Distinctive Features and Their Correlates: M.I.T. Press.
Jamison HL, Watkins KE, Bishop DVM, Matthews PM (2006) Hemispheric Specialization for Processing Auditory Nonspeech Stimuli. Cerebral Cortex 16:1266-1275.
Jehee JFM, Ballard DH (2009) Predictive Feedback Can Account for Biphasic Responses in the Lateral Geniculate Nucleus. PLOS Computational Biology 5:e1000373.
Johnson S, Prendergast G, Hymers M, Green G (2011) Examining the Effects of One- and Three-Dimensional Spatial Filtering Analyses in Magnetoencephalography. PLOS ONE 6:e22251.
Jones G, Macken B (2015) Questioning short-term memory and its measurement: Why digit span measures long-term associative learning. Cognition 144:1-13.
Joris PX, Schreiner CE, Rees A (2004) Neural Processing of Amplitude-Modulated Sounds. Physiological Reviews 84:541.
Junghofer M, Elbert T, Tucker DM, Rockstroh B (2000) Statistical control of artifacts in dense array EEG/MEG studies. Psychophysiology 37:523-532.
Kastner S, Pinsk MA, De Weerd P, Desimone R, Ungerleider LG (1999) Increased activity in human visual cortex during directed attention in the absence of visual stimulation. Neuron 22:751-761.
Kaufman AS, Kaufman NL (2004) Kaufman brief intelligence test: Wiley Online Library.
Keitel A, Ince RAA, Gross J, Kayser C (2017) Auditory cortical delta-entrainment interacts with oscillatory power in multiple fronto-parietal networks. NeuroImage 147:32-42.
Kemper S, Anagnopoulos C (2008) Language and Aging. Annual Review of Applied Linguistics 10:37-50.
Kerlin JR, Shahin AJ, Miller LM (2010) Attentional Gain Control of Ongoing Cortical Speech Representations in a "Cocktail Party". Journal of Neuroscience 30:620-628.
Keshner MS (1982) 1/f noise. Proceedings of the IEEE 70:212-218.
Khalighinejad B, Cruzatto da Silva G, Mesgarani N (2017) Dynamic Encoding of Acoustic Features in Neural Responses to Continuous Speech. The Journal of Neuroscience.
Kim JJ, Crespo-Facorro B, Andreasen NC, O'Leary DS, Zhang B, Harris G, Magnotta VA (2000) An MRI-based parcellation method for the temporal lobe. Neuroimage 11:271-288.
Knill DC, Pouget A (2004) The Bayesian brain: the role of uncertainty in neural coding and computation. Trends in neurosciences 27:712-719.
Kok P, Jehee Janneke FM, de Lange Floris P (2012a) Less Is More: Expectation Sharpens Representations in the Primary Visual Cortex. Neuron 75:265-270.

Kok P, Rahnev D, Jehee JF, Lau HC, de Lange FP (2012b) Attention reverses the effect of prediction in silencing sensory signals. Cerebral cortex (New York, NY : 1991) 22:2197-2206.
Korzeniewska A, Manczak M, Kaminski M, Blinowska KJ, Kasicki S (2003) Determination of information flow direction among brain structures by a modified directed transfer function (dDTF) method. Journal of neuroscience methods 125:195-207.
Kösem A, van Wassenhove V (2016) Distinct contributions of low- and high-frequency neural oscillations to speech comprehension. Language, Cognition and Neuroscience:1-9.
Kozinska D, Carducci F, Nowinski K (2001) Automatic alignment of EEG/MEG and MRI data sets. Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology 112:1553-1561.
Kruskal JB, Wish M (1978) Multidimensional Scaling: SAGE Publications.
Kuhl PK (2004) Early language acquisition: cracking the speech code. Nature reviews Neuroscience 5:831-843.
Kuhl PK (2010) Brain mechanisms in early language acquisition. Neuron 67:713-727.
Kuhl PK, Coffey-Corina S, Padden D, Dawson G (2005) Links between social and linguistic processing of speech in preschool children with autism: behavioral and electrophysiological measures. Developmental science 8:F1-F12.
Kus R, Kaminski M, Blinowska KJ (2004) Determination of EEG activity propagation: pair-wise versus multichannel estimate. IEEE Trans Biomed Eng 51:1501-1510.
Kuwabara H (1996) Acoustic properties of phonemes in continuous speech for different speaking rate. In: Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on, pp 2435-2438 vol.2434.
Lahav A, Skoe E (2014) An acoustic gap between the NICU and womb: a potential risk for compromised neuroplasticity of the auditory system in preterm infants. Frontiers in Neuroscience 8.
Lalor EC, Foxe JJ (2010) Neural responses to uninterrupted natural speech can be extracted with precise temporal resolution. European Journal of Neuroscience 31:189-193.
Lalor EC, Power AJ, Reilly RB, Foxe JJ (2009) Resolving Precise Temporal Processing Properties of the Auditory System Using Continuous Stimuli. J Neurophysiol 102:349-359.
Lalor EC, Pearlmutter BA, Reilly RB, McDarby G, Foxe JJ (2006) The VESPA: a method for the rapid estimation of a visual evoked potential. NeuroImage 32:1549-1561.
Lau EF, Phillips C, Poeppel D (2008) A cortical network for semantics: (de)constructing the N400. Nature reviews Neuroscience 9:920-933.
Lee TS, Mumford D (2003) Hierarchical Bayesian inference in the visual cortex. Journal of the Optical Society of America A, Optics, image science, and vision 20:1434-1448.
Lehmann D, Skrandies W (1980) Reference-free identification of components of checkerboard-evoked multichannel potential fields. Electroencephalography and clinical neurophysiology 48:609-621.
Lehongre K, Morillon B, Giraud A-L, Ramus F (2013) Impaired auditory sampling in dyslexia: further evidence from combined fMRI and EEG.
Lehongre K, Ramus F, Villiermet N, Schwartz D, Giraud AL (2011) Altered low-gamma sampling in auditory cortex accounts for the three main facets of dyslexia. Neuron 72:1080-1090.
Leonard LB (2014) Children with Specific Language Impairment: MIT Press.

Leonard MK, Chang EF (2014) Dynamic speech representations in the human temporal lobe. Trends Cogn Sci 18:472-479.
Leonard MK, Baud MO, Sjerps MJ, Chang EF (2016) Perceptual restoration of masked speech in human cortex. Nature Communications 7:13619.
Lewis AG, Bastiaansen M (2015) A predictive coding framework for rapid neural dynamics during sentence-level language comprehension. Cortex 68:155-168.
Li X, Branch CA, DeLisi LE (2009) Language pathway abnormalities in schizophrenia: a review of fMRI and other imaging studies. Current opinion in psychiatry 22:131-139.
Liberman A (1970) Some characteristics of perception in the speech mode. Perception and its Disorders 48:238-254.
Liberman AM, Cooper FS, Shankweiler DP, Studdert-Kennedy M (1967) Perception of the speech code. Psychological review 74:431.
Liebenthal E, Binder JR, Spitzer SM, Possing ET, Medler DA (2005) Neural Substrates of Phonemic Perception. Cerebral Cortex 15:1621-1631.
Logothetis NK (2008) What we can do and what we cannot do with fMRI. Nature 453:869-878.
Long Michael A, Katlowitz Kalman A, Svirsky Mario A, Clary Rachel C, Byun Tara M, Majaj N, Oya H, Howard Iii Matthew A, Greenlee Jeremy DW (2016) Functional Segregation of Cortical Regions Underlying Speech Timing and Articulation. Neuron 89:1187-1193.
Lopez-Valdes A, McLaughlin M, Viani L, Walshe P, Smith J, Zeng F-G, Reilly RB (2013) Auditory mismatch negativity in cochlear implant users: A window to spectral discrimination. In: Engineering in Medicine and Biology Society (EMBC), 2013 35th Annual International Conference of the IEEE, pp 3555-3558: IEEE.
Luck SJ (2005) An introduction to the event-related potential technique.
Ludmer R, Dudai Y, Rubin N (2011) Uncovering Camouflage: Amygdala Activation Predicts Long-Term Memory of Induced Perceptual Insight. Neuron 69:1002-1014.
Luo H, Poeppel D (2007) Phase patterns of neuronal responses reliably discriminate speech in human auditory cortex. Neuron 54:1001-1010.
Luo H, Liu Z, Poeppel D (2010) Auditory cortex tracks both auditory and visual stimulus dynamics using low-frequency neuronal phase modulation. PLoS Biol 8:e1000445.
Lütkepohl H (2007) New Introduction to Multiple Time Series Analysis: Springer Berlin Heidelberg.
Macleod A, Summerfield Q (1987) Quantifying the contribution of vision to speech perception in noise. British Journal of Audiology 21:131-141.
MacQueen J (1967) Some methods for classification and analysis of multivariate observations. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, pp 281-297. Berkeley, Calif.: University of California Press.
Makeig S, Jung T-P, Ghahremani D, Sejnowski TJ (1996) Independent component analysis of simulated ERP data. Institute for Neural Computation, University of California: technical report INC-9606.
Maris E, Oostenveld R (2007) Nonparametric statistical testing of EEG-and MEG-data. Journal of neuroscience methods 164:177-190.
Mason RA, Just MA (2007) Lexical ambiguity in sentence comprehension. Brain research 1146:115-127.

Massaro DW (1989) Testing between the TRACE model and the fuzzy logical model of speech perception. Cognitive psychology 21:398-421.
McClelland JL (2013) Integrating probabilistic models of perception and interactive neural networks: a historical and tutorial review. Front Psychol 4:503.
McClelland JL, Elman JL (1986) The TRACE model of speech perception. Cognitive psychology 18:1-86.
McClelland JL, Mirman D, Holt LL (2006) Are there interactive processes in speech perception? Trends Cogn Sci 10:363-369.
McGettigan C, Evans S, Rosen S, Agnew ZK, Shah P, Scott SK (2012) An application of univariate and multivariate approaches in FMRI to quantifying the hemispheric lateralization of acoustic and linguistic processes. J Cogn Neurosci 24:636-652.
McNealy K, Mazziotta JC, Dapretto M (2006) Cracking the Language Code: Neural Mechanisms Underlying Speech Parsing. The Journal of neuroscience : the official journal of the Society for Neuroscience 26:7629-7639.
McNulty MA (2003) Dyslexia and the Life Course. Journal of learning disabilities 36:363-381.
Mesgarani N, Chang EF (2012) Selective cortical representation of attended speaker in multi-talker speech perception. Nature 485:233-U118.
Mesgarani N, David SV, Fritz JB, Shamma SA (2008) Phoneme representation and classification in primary auditory cortex. J Acoust Soc Am 123:899-909.
Mesgarani N, David SV, Fritz JB, Shamma SA (2009) Influence of Context and Behavior on Stimulus Reconstruction From Neural Activity in Primary Auditory Cortex. J Neurophysiol 102:3329-3339.
Mesgarani N, Cheung C, Johnson K, Chang EF (2014) Phonetic Feature Encoding in Human Superior Temporal Gyrus. Science 343:1006-1010.
Mesulam MM, Rogalski EJ, Wieneke C, Hurley RS, Geula C, Bigio EH, Thompson CK, Weintraub S (2014) Primary progressive aphasia and the evolving neurology of the language network. Nat Rev Neurol 10:554-569.
Meyer T, Ramachandran S, Olson CR (2014) Statistical learning of serial visual transitions by neurons in monkey inferotemporal cortex. J Neurosci 34:9332-9337.
Millman RE, Johnson SR, Prendergast G (2015) The role of phase-locking to the temporal envelope of speech in auditory perception and speech intelligibility. Journal of cognitive neuroscience 27:533-545.
Millman RE, Prendergast G, Hymers M, Green GG (2013) Representations of the temporal envelope of sounds in human auditory cortex: can the results from invasive intracortical "depth" electrode recordings be replicated using non-invasive MEG "virtual electrodes"? NeuroImage 64:185-196.
Mirkovic B, Debener S, Jaeger M, De Vos M (2015) Decoding the attended speech stream with multi-channel EEG: implications for online, daily-life applications. J Neural Eng 12:046007.
Mirman D, McClelland JL, Holt LL (2006) An interactive Hebbian account of lexically guided tuning of speech perception. Psychonomic bulletin & review 13:958-965.
Mody M, Belliveau JW (2013) Speech and language impairments in autism: insights from behavior and neuroimaging. North American journal of medicine & science 5:157.
Moerel M, De Martino F, Formisano E (2014) An anatomical and functional topography of human auditory cortical areas. Frontiers in Neuroscience 8:225.

Moore BCJ, Glasberg BR (1983) Suggested formulae for calculating auditory-filter bandwidths and excitation patterns. The Journal of the Acoustical Society of America 74:750-753.
Morillon B, Lehongre K, Frackowiak RS, Ducorps A, Kleinschmidt A, Poeppel D, Giraud AL (2010) Neurophysiological origin of human brain asymmetry for speech and language. Proceedings of the National Academy of Sciences of the United States of America 107:18688-18693.
Morosan P, Rademacher J, Schleicher A, Amunts K, Schormann T, Zilles K (2001) Human primary auditory cortex: cytoarchitectonic subdivisions and mapping into a spatial reference system. Neuroimage 13:684-701.
Mumford D (1992) On the computational architecture of the neocortex. II. The role of cortico-cortical loops. Biological cybernetics 66:241-251.
Murray SO, Kersten D, Olshausen BA, Schrater P, Woods DL (2002) Shape perception reduces activity in human primary visual cortex. Proceedings of the National Academy of Sciences of the United States of America 99:15164-15169.
Naatanen R, Winkler I (1999) The concept of auditory stimulus representation in cognitive neuroscience. Psychol Bull 125:826-859.
Norris D, McQueen JM, Cutler A (2000) Merging information in speech recognition: feedback is never necessary. The Behavioral and brain sciences 23:299-325; discussion 325-270.
Norris D, McQueen JM, Cutler A (2003) Perceptual learning in speech. Cognitive psychology 47:204-238.
Norris D, McQueen JM, Cutler A (2016) Prediction, Bayesian inference and feedback in speech recognition. Lang Cogn Neurosci 31:4-18.
Nourski KV (2017) Auditory processing in the human cortex: An intracranial electrophysiology perspective. Laryngoscope Investigative Otolaryngology.
Nourski KV, Reale RA, Oya H, Kawasaki H, Kovach CK, Chen H, Howard MA, 3rd, Brugge JF (2009) Temporal envelope of time-compressed speech represented in the human auditory cortex. The Journal of neuroscience : the official journal of the Society for Neuroscience 29:15564-15574.
O'Sullivan JA, Power AJ, Mesgarani N, Rajaram S, Foxe JJ, Shinn-Cunningham BG, Slaney M, Shamma SA, Lalor EC (2014) Attentional Selection in a Cocktail Party Environment Can Be Decoded from Single-Trial EEG. Cerebral Cortex:bht355.
O'Sullivan AE, Crosse MJ, Di Liberto GM, Lalor EC (2016) Visual Cortical Entrainment to Motion and Categorical Speech Features during Silent Lipreading. Frontiers in Human Neuroscience 10:679.
Obleser J, Kotz SA (2009) Expectancy constraints in degraded speech modulate the language comprehension network. Cerebral Cortex:bhp128.
Obleser J, Eisner F, Kotz SA (2008) Bilateral Speech Comprehension Reflects Differential Sensitivity to Spectral and Temporal Features. The Journal of Neuroscience 28:8116-8123.
Obleser J, Herrmann B, Henry MJ (2012) Neural Oscillations in Speech: Don't be Enslaved by the Envelope. Frontiers in Human Neuroscience 6:250.
Obleser J, Zimmermann J, Van Meter J, Rauschecker JP (2007) Multiple stages of auditory speech perception reflected in event-related FMRI. Cerebral Cortex 17:2251-2257.
Okada K, Rong F, Venezia J, Matchin W, Hsieh IH, Saberi K, Serences JT, Hickok G (2010) Hierarchical organization of human auditory cortex: evidence from acoustic invariance in the response to intelligible speech. Cerebral cortex (New York, NY : 1991) 20:2486-2495.

Overath T, McDermott JH, Zarate JM, Poeppel D (2015) The cortical analysis of speech-specific temporal structure revealed by responses to sound quilts. Nat Neurosci 18:903-911.
Park H, Ince RA, Schyns PG, Thut G, Gross J (2015) Frontal top-down signals increase coupling of auditory low-frequency oscillations to continuous speech in human listeners. Curr Biol 25:1649-1653.
Pasley BN, David SV, Mesgarani N, Flinker A, Shamma SA, Crone NE, Knight RT, Chang EF (2012) Reconstructing Speech from Human Auditory Cortex. PLoS Biol 10.
Paul R (2007) Language Disorders from Infancy Through Adolescence: Assessment & Intervention: Mosby Elsevier.
Paulraj MP, Subramaniam K, Yaccob SB, Adom AHB, Hema CR (2015) Auditory Evoked Potential Response and Hearing Loss: A Review. Open Biomed Eng J 9:17-24.
Peelle JE (2012) The hemispheric lateralization of speech processing depends on what "speech" is: a hierarchical perspective. Frontiers in Human Neuroscience 6.
Peelle JE, Johnsrude IS, Davis MH (2010) Hierarchical Processing for Speech in Human Auditory Cortex and Beyond. Frontiers in Human Neuroscience 4:51.
Peelle JE, Gross J, Davis MH (2013) Phase-locked responses to speech in human auditory cortex are enhanced during comprehension. Cerebral cortex (New York, NY : 1991) 23:1378-1387.
Peter V, Kalashnikova M, Burnham D (2016) Neural processing of amplitude and formant rise time in dyslexia. Developmental Cognitive Neuroscience 19:152-163.
Picton T (2013) Hearing in time: evoked potential studies of temporal processing. Ear Hear 34:385-401.
Picton TW, Hillyard SA, Krausz HI, Galambos R (1974) Human auditory evoked potentials. I: Evaluation of components. Electroencephalography and clinical neurophysiology 36:179-190.
Pitt MA, Samuel AG (1993) An empirical and meta-analytic evaluation of the phoneme identification task. Journal of Experimental Psychology: Human Perception and Performance 19:699-725.
Poelmans H, Luts H, Vandermosten M, Boets B, Ghesquiere P, Wouters J (2012) Auditory steady state cortical responses indicate deviant phonemic-rate processing in adults with dyslexia. Ear Hear 33:134-143.
Poeppel D (2003) The analysis of speech in different temporal integration windows: cerebral lateralization as 'asymmetric sampling in time'. Speech communication 41:245-255.
Poeppel D (2006) Language: Specifying the Site of Modality-Independent Meaning. Current Biology 16:R930-R932.
Poeppel D (2014) The neuroanatomic and neurophysiological infrastructure for speech and language. Current Opinion in Neurobiology 28:142-149.
Power AJ, Lalor EC, Reilly RB (2011) Endogenous Auditory Spatial Attention Modulates Obligatory Sensory Activity in Auditory Cortex. Cerebral Cortex 21:1223-1230.
Power AJ, Foxe JJ, Forde EJ, Reilly RB, Lalor EC (2012) At what time is the cocktail party? A late locus of selective attention to natural speech. European Journal of Neuroscience 35:1497-1503.
Purves D, Augustine G, Fitzpatrick D, Hall W, LaMantia A, McNamara J, White L (2008) Neuroscience. Sunderland, MA, USA: Sinauer Associates, Inc.

Rademacher J, Morosan P, Schormann T, Schleicher A, Werner C, Freund HJ, Zilles K (2001) Probabilistic mapping and volume measurement of human primary auditory cortex. NeuroImage 13:669-683.
Rao RP, Ballard DH (1999) Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat Neurosci 2:79-87.
Rauschecker JP, Scott SK (2009) Maps and streams in the auditory cortex: nonhuman primates illuminate human speech processing. Nat Neurosci 12:718-724.
Reed V (2012) An Introduction to Children with Language Disorders: Pearson.
Richlan F (2012) Developmental dyslexia: dysfunction of a left hemisphere reading network. Frontiers in human neuroscience 6:120.
Rijsbergen CJV (1979) Information Retrieval: Butterworth-Heinemann.
Rivier F, Clarke S (1997) Cytochrome oxidase, acetylcholinesterase, and NADPH-diaphorase staining in human supratemporal and insular cortex: evidence for multiple auditory areas. Neuroimage 6:288-304.
Ross LA, Saint-Amour D, Leavitt VM, Molholm S, Javitt DC, Foxe JJ (2007) Impaired multisensory processing in schizophrenia: Deficits in the visual enhancement of speech comprehension under noisy environmental conditions. Schizophrenia Research 97:173–183.
Rubin N, Nakayama K, Shapley R (1997) Abrupt learning and retinal size specificity in illusory-contour perception. Curr Biol 7:461-467.
Ruggles D, Bharadwaj H, Shinn-Cunningham BG (2012) Why middle-aged listeners have trouble hearing in everyday settings. Curr Biol 22:1417-1422.
Ruhnau P, Herrmann B, Maess B, Brauer J, Friederici A, Schröger E (2013) Processing of complex distracting sounds in school-aged children and adults: evidence from EEG and MEG data. Frontiers in Psychology 4.
Sabornie EJ (1994) Social-Affective Characteristics in Early Adolescents Identified as Learning Disabled and Nondisabled. Learning Disability Quarterly 17:268-279.
Saenz M, Langers DR (2014) Tonotopic mapping of human auditory cortex. Hear Res 307:42-52.
Salmelin R (2007) Clinical neurophysiology of language: the MEG approach. Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology 118:237-254.
Sasaki Y (2007) The truth of the F-measure. Teach Tutor mater 1.
Schnupp J, Nelken I, King A (2011) Auditory Neuroscience: Making Sense of Sound: MIT Press.
Schonwiesner M, Rubsamen R, von Cramon DY (2005) Hemispheric asymmetry for spectral and temporal processing in the human antero-lateral auditory belt cortex. The European journal of neuroscience 22:1521-1528.
Schonwiesner M, Novitski N, Pakarinen S, Carlson S, Tervaniemi M, Naatanen R (2007) Heschl's gyrus, posterior superior temporal gyrus, and mid-ventrolateral prefrontal cortex have different roles in the detection of acoustic changes. J Neurophysiol 97:2075-2082.
Schroeder CE, Lakatos P (2009) Low-frequency neuronal oscillations as instruments of sensory selection. Trends in neurosciences 32:9-18.
Scott SK, Johnsrude IS (2003) The neuroanatomical and functional organization of speech perception. Trends in neurosciences 26:100-107.
Sedley W, Gander PE, Kumar S, Kovach CK, Oya H, Kawasaki H, Howard MA, III, Griffiths TD (2016) Neural signatures of perceptual inference. eLife 5:e11476.

Semel EM, Wiig EH, Secord W (2006) Clinical evaluation of language fundamentals: Pearson Assessment.
Shalom DB, Poeppel D (2008) Functional anatomic models of language: assembling the pieces. The Neuroscientist : a review journal bringing neurobiology, neurology and psychiatry 14:119-127.
Shamma S (2014) How phonetically selective is the human auditory cortex? Trends in Cognitive Sciences 18:391-392.
Shannon RV, Zeng FG, Kamath V, Wygonski J, Ekelid M (1995) Speech recognition with primarily temporal cues. Science 270:303-304.
Shepard RN (1980) Multidimensional scaling, tree-fitting, and clustering. Science 210:390-398.
Simanova I, Hagoort P, Oostenveld R, van Gerven MAJ (2014) Modality-Independent Decoding of Semantic Information from the Human Brain. Cerebral Cortex 24:426-434.
Snowling MJ (2000) Dyslexia.
Sohoglu E, Davis MH (2016) Perceptual learning of degraded speech by minimizing prediction error. Proceedings of the National Academy of Sciences of the United States of America 113:E1747-1756.
Sohoglu E, Peelle JE, Carlyon RP, Davis MH (2012) Predictive top-down integration of prior knowledge during speech perception. The Journal of neuroscience : the official journal of the Society for Neuroscience 32:8443-8453.
Spratling MW (2008) Reconciling Predictive Coding and Biased Competition Models of Cortical Function. Frontiers in Computational Neuroscience 2:4.
Stanley GB, Li FF, Dan Y (1999) Reconstruction of natural scenes from ensemble responses in the lateral geniculate nucleus. Journal of Neuroscience 19:8036-8042.
Steinschneider M, Nourski KV, Fishman YI (2013) Representation of speech in human auditory cortex: is it special? Hear Res 305:57-73.
Stephan KE, Harrison LM, Kiebel SJ, David O, Penny WD, Friston KJ (2007) Dynamic causal models of neural system dynamics:current state and future extensions. J Biosci 32:129-144.
Stevens KN, House AS (1972) Speech perception(Acoustic model and linguistic, syntactic, lexical and semantic factors in speech perception and production process). Foundations of modern auditory theory 2:3-62.
Summerfield C, Koechlin E (2008) A Neural Representation of Prior Information during Perceptual Inference. Neuron 59:336-347.
Summerfield C, Trittschuh EH, Monti JM, Mesulam MM, Egner T (2008) Neural repetition suppression reflects fulfilled perceptual expectations. Nat Neurosci 11:1004-1006.
Summerfield C, Egner T, Greene M, Koechlin E, Mangels J, Hirsch J (2006) Predictive Codes for Forthcoming Perception in the Frontal Cortex. Science 314:1311-1314.
Swanson JM (1992) School-based assessments and interventions for ADD students: KC publishing.
Telkemeyer S, Rossi S, Koch SP, Nierhaus T, Steinbrink J, Poeppel D, Obrig H, Wartenburger I (2009) Sensitivity of newborn auditory cortex to the temporal structure of sounds. J Neurosci 29:14726-14733.
Theunissen FE, David SV, Singh NC, Hsu A, Vinje WE, Gallant JL (2001) Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. Network: Computation in Neural Systems 12:289-316.

Thomson JM, Goswami U (2008) Rhythmic processing in children with developmental dyslexia: auditory and motor rhythms link to reading and spelling. Journal of physiology, Paris 102:120-129.
Thwaites A, Nimmo-Smith I, Fonteneau E, Patterson RD, Buttery P, Marslen-Wilson WD (2015) Tracking cortical entrainment in neural activity: auditory processes in human temporal cortex. Frontiers in Computational Neuroscience 9.
Tikhonov AN, Arsenin VIAk, John F (1977) Solutions of ill-posed problems: Winston Washington, DC.
Todorovic A, van Ede F, Maris E, de Lange FP (2011) Prior expectation mediates neural adaptation to repeated sounds in the auditory cortex: an MEG study. J Neurosci 31:9118-9123.
Tomblin JB, Records NL, Zhang X (1996) A System for the Diagnosis of Specific Language Impairment in Kindergarten Children. Journal of Speech, Language, and Hearing Research 39:1284-1294.
Torgesen JK, Wagner RK, Rashotte CA (2012) TOWRE: Test of word reading efficiency: Psychological Corporation.
Tuennerhoff J, Noppeney U (2016) When sentences live up to your expectations. NeuroImage 124:641-653.
Turkeltaub PE, Coslett HB (2010) Localization of sublexical speech perception components. Brain and language 114:1-15.
Turken AU, Dronkers NF (2011) The neural architecture of the language comprehension network: converging evidence from lesion and connectivity analyses. Frontiers in systems neuroscience 5:1.
van Ede F, Jensen O, Maris E (2010) Tactile expectation modulates pre-stimulus β-band oscillations in human sensorimotor cortex. NeuroImage 51:867-876.
Van Veen BD, van Drongelen W, Yuchtman M, Suzuki A (1997) Localization of brain electrical activity via linearly constrained minimum variance spatial filtering. IEEE Trans Biomed Eng 44:867-880.
Vellutino FR, Fletcher JM, Snowling MJ, Scanlon DM (2004) Specific reading disability (dyslexia): what have we learned in the past four decades? Journal of child psychology and psychiatry, and allied disciplines 45:2-40.
Wagner RK, Torgesen JK, Rashotte CA, Pearson NA (2013) Comprehensive Test of Phonological Processing: CTOPP2.
Wang X, Lu T, Liang L (2003) Cortical processing of temporal modulations. Speech Communication 41:107-121.
Warren RM (1970) Perceptual restoration of missing speech sounds. Science 167:392-393.
Wechsler D (2009) Wechsler Individual Achievement Test (WI AT - III). The Psychological Corporation, TX, US.
Wernicke C (1874) Der aphasische Symptomencomplex, eine psychologische Studie auf anatomischer Basis, von Dr. C. Wernicke. Breslau: M. Cohn und Weigert.
Wiener J, Schneider BH (2002) A multisource exploration of the friendship patterns of children with and without learning disabilities. Journal of abnormal child psychology 30:127-141.
Wild CJ, Davis MH, Johnsrude IS (2012) Human auditory cortex is sensitive to the perceived clarity of speech. NeuroImage 60:1490-1502.
Willems G, Jansma B, Blomert L, Vaessen A (2016) Cognitive and familial risk evidence converged: A data-driven identification of distinct and homogeneous subtypes within the heterogeneous sample of reading disabled children. Research in Developmental Disabilities 53–54:213-231.

Williams RW, Herrup K (1988) The control of neuron number. Annu Rev Neurosci 11:423-453.
Wornell GW (1993) Wavelet-based representations for the 1/<e1>f</e1> family of fractal processes. Proceedings of the IEEE 81:1428-1450.
Yang X, Wang K, Shamma SA (1992) Auditory representations of acoustic signals. IEEE Transactions on Information Theory 38:824-839.
Yuan J, Liberman M (2008) Speaker identification on the SCOTUS corpus. Journal of the Acoustical Society of America 123:3878.
Yuille A, Kersten D (2006) Vision as Bayesian inference: analysis by synthesis? Trends in Cognitive Sciences 10:301-308.
Zatorre R, Evans A, Meyer E, Gjedde A (1992) Lateralization of phonetic and pitch discrimination in speech processing. Science 256:846-849.
Zatorre RJ, Belin P (2001) Spectral and temporal processing in human auditory cortex. Cerebral cortex 11:946-953.
Zatorre RJ, Salimpoor VN (2013) From perception to pleasure: Music and its neural substrates. Proceedings of the National Academy of Sciences of the United States of America 110:10430-10437.
Zatorre RJ, Belin P, Penhune VB (2002) Structure and function of auditory cortex: music and speech. Trends Cogn Sci 6:37-46.
Zelano C, Mohanty A, Gottfried Jay A (2011) Olfactory Predictive Codes and Stimulus Templates in Piriform Cortex. Neuron 72:178-187.
Zhang L, Yue Q, Zhang Y, Shu H, Li P (2015) Task-dependent modulation of regions in the left temporal cortex during auditory sentence comprehension. Neuroscience Letters 584:351-355.
Zion Golumbic EM, Cogan GB, Schroeder CE, Poeppel D (2013a) Visual input enhances selective speech envelope tracking in auditory cortex at a “cocktail party”. The Journal of Neuroscience 33:1417-1426.
Zion Golumbic EM, Ding N, Bickel S, Lakatos P, Schevon CA, McKhann GM, Goodman RR, Emerson R, Mehta AD, Simon JZ (2013b) Mechanisms Underlying Selective Neuronal Tracking of Attended Speech at a “Cocktail Party”. Neuron 77:980-991.
Zoefel B, VanRullen R (2015) Selective perceptual phase entrainment to speech rhythm in the absence of spectral energy fluctuations. The Journal of neuroscience : the official journal of the Society for Neuroscience 35:1954-1964.
Zoefel B, VanRullen R (2016) EEG oscillations entrain their phase to high-level features of speech sound. NeuroImage 124:16-23.